

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2002-082775

(43)Date of publication of application : 22.03.2002

(51)Int.Cl.

G06F 3/06

(21)Application number : 2001-053458

(71)Applicant : HITACHI LTD

(22)Date of filing : 28.02.2001

(72)Inventor : ARAKAWA TAKASHI
MOGI KAZUHIKO
EGUCHI KENTETSU
ARAI HIROHARU

(30)Priority

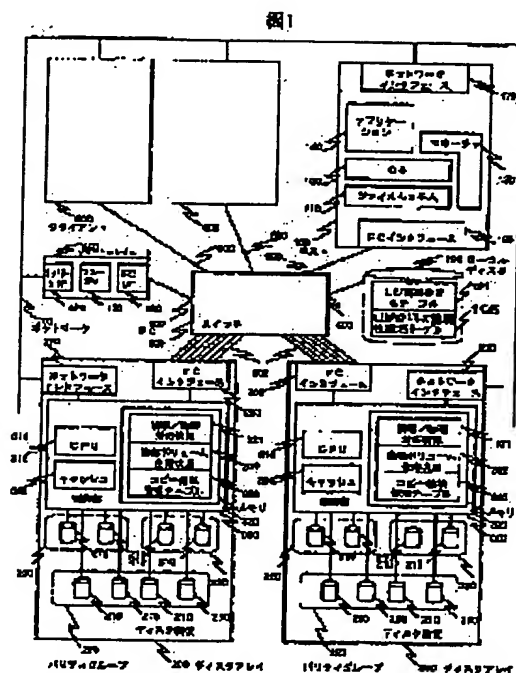
Priority number : 2000205510 Priority date : 06.07.2000 Priority country : JP

(54) COMPUTER SYSTEM

(57)Abstract:

PROBLEM TO BE SOLVED: To properly perform relocation of data among different storage sub- systems.

SOLUTION: A disk array 200 acquires a use state of a disk device 210 for read/write from a host 100. The host 100 decides an LU of a relocating destination of the data stored in a relocating object LU by collecting use states from plural disk arrays 200. Then a logical position name table 191 to decide correspondence between a logical position name of the data as a data position for application and the LU is changed. In addition, the relocation of data among the different disk arrays 200 is performed by transferring the data stored in the LU being the relocating object to an LU at the relocating destination among the disk arrays 200.



LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

Copyright (C); 1998,2003 Japan Patent Office

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号
特開2002-82775
(P2002-82775A)

(43) 公開日 平成14年3月22日 (2002.3.22)

(51) Int.Cl.⁷

G 0 6 F 3/06

識別記号

3 0 2

5 4 0

F I

G 0 6 F 3/06

ターマコト* (参考)

3 0 2 J 5 B 0 6 5

5 4 0

審査請求 未請求 請求項の数14 O L (全 39 頁)

(21) 出願番号 特願2001-53458(P2001-53458)
(22) 出願日 平成13年2月28日 (2001.2.28)
(31) 優先権主張番号 特願2000-205510(P2000-205510)
(32) 優先日 平成12年7月6日 (2000.7.6)
(33) 優先権主張国 日本 (J P)

(71) 出願人 000005108
株式会社日立製作所
東京都千代田区神田駿河台四丁目6番地
(72) 発明者 荒川 敬史
神奈川県川崎市麻生区王禅寺1099番地 株
式会社日立製作所システム開発研究所内
(72) 発明者 茂木 和彦
神奈川県川崎市麻生区王禅寺1099番地 株
式会社日立製作所システム開発研究所内
(74) 代理人 100075096
弁理士 作田 康夫

最終頁に続く

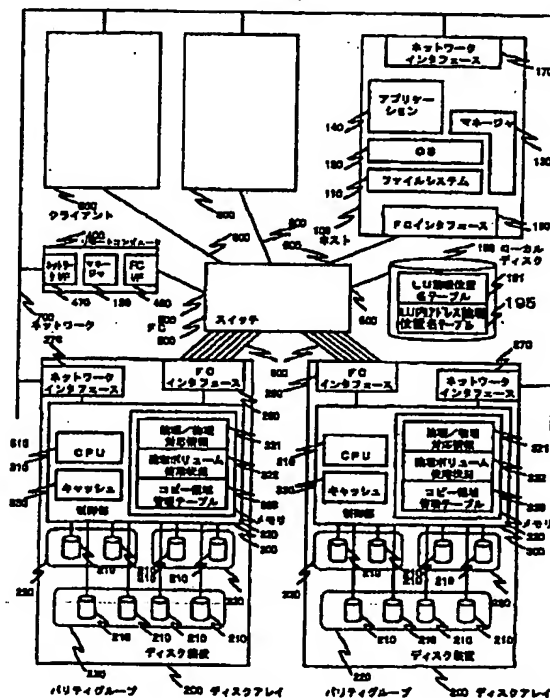
(54) 【発明の名称】 計算機システム

(57) 【要約】

【課題】 異なるストレージサブシステム間でのデータの再配置を適正に行えるようにする。

【解決手段】 ディスクアレイ200は、ホスト100からのリード/ライトに対してディスク装置210の使用状況を取得する。ホスト100は、複数のディスクアレイ200から使用状況を収集して、再配置対象LUに格納されているデータの再配置先のLUを決定する。そして、アプリケーションにとってのデータ位置であるデータの論理位置名とLUとの対応を決定するLU論理位置名テーブル191を変更する。また、ディスクアレイ200間で、再配置対象となったLUに格納されているデータを、再配置先のLUに移動することで、異なるディスクアレイ200間でのデータの再配置を行う。

図1



【特許請求の範囲】

【請求項1】 計算機と、前記計算機に接続される複数のストレージサブシステムとを有し、

前記複数のストレージサブシステムの各々は、

該ストレージサブシステムが有する記憶領域の各々について、前記記憶領域の使用状況を取得する取得手段を有し、

前記計算機は、

前記複数のストレージサブシステムの各々から、前記取得手段によって取得された記憶領域各々の使用状況を集める収集手段を有することを特徴とする計算機システム。

【請求項2】 前記記憶領域は、ディスク装置を用いて形成される論理的な記憶領域であることを特徴とする請求項1記載の計算機システム。

【請求項3】 前記複数のストレージサブシステムの各々は、前記計算機の指示にしたがって、他の前記複数のストレージサブシステムのうちのいずれかからデータの移動を行う移動手段と、前記移動手段が完了したことを前記計算機に通知する手段とを有し、

前記計算機は、

前記データの論理的な位置と、前記ストレージサブシステムにおいて前記データが格納される記憶領域との対応を記憶する対応記憶部と、

前記通知手段によって前記ストレージサブシステム間でのデータの移動が完了した通知を受け取った場合に、前記対応記憶部を更新する更新手段とを有することを特徴とする請求項2記載の計算機システム。

【請求項4】 前記計算機は、前記複数のストレージサブシステム各々から収集された各々の前記記憶領域の使用状況を表示する表示部と、

該計算機システムの使用者に、前記表示部に表示された各々の前記記憶領域のいずれかを選択させる選択手段とを有し、

前記指示には、前記選択手段によって選択された記憶領域を指定する情報が含まれていることを特徴とする請求項3記載の計算機システム。

【請求項5】 前記計算機は、前記複数のストレージサブシステム各々から収集された各々の前記記憶領域の使用状況に基づいて、所定の条件を満たす前記記憶領域を選択する手段を有し、前記指示には、前記選択手段によって選択された記憶領域を指定する情報が含まれていることを特徴とする請求項3記載の計算機システム。

【請求項6】 前記所定の条件とは、前記記憶領域へのアクセスの頻度が一定値以上であるという条件であることを特徴とする請求項5記載の計算機システム。

【請求項7】 前記計算機は、前記データと前記計算機で実行されるファイルシステムが管理するファイルとの対応関係を保持するファイル対応関係記録部と、前記ファイル対応関係記録部を使用して前記ファイルご

とに前記記憶領域の使用状況を収集する手段とを有し、前記複数のストレージサブシステムの各々は、前記指示にしたがって、指定された前記ファイルを他の前記複数のストレージサブシステムのうちのいずれかから移動する手段と前記移動する手段が完了したことを前記計算機に通知する手段を有することを特徴とする請求項3記載の計算機システム。

【請求項8】 前記計算機は、前記通知する手段によって、ストレージサブシステム間のファイルの移動が完了したことを通知されると、前記ファイル対応関係記録部を更新する手段を有することを特徴とする請求項7記載の計算機システム。

【請求項9】 前記使用状況には、データを読み出す際に前記記憶領域が占有される時間が含まれることを特徴とする、請求項1～8記載の計算機システム。

【請求項10】 ホストコンピュータと接続される接続部と、複数の記憶装置と、

前記複数の記憶装置を制御する制御部と、前記複数の記憶装置の使用状況を示す情報が格納される記憶部とを有し、

前記制御部は、前記接続部を介して、前記複数の記憶装置の使用状況を示す情報の転送を要求する命令を受け取った場合には、前記記憶部から前記複数の記憶装置の使用状況を示す情報を読み出し、前記ホストコンピュータに転送することを特徴とするストレージサブシステム。

【請求項11】 複数のストレージサブシステムと接続される接続部と、前記複数のストレージサブシステムから取得した前記複数のストレージサブシステム各々が有する記憶領域の使用状況が格納される記憶部と、中央演算処理装置とを有し、

前記中央演算処理装置は、前記記憶部に格納された前記複数のストレージサブシステムの使用状況の情報に基づいて、移動すべきデータを決定する決定手段と、

前記決定手段によって決定されたデータの情報を前記接続部を介して、移動先となる前記複数のストレージサブシステムのうちのいずれか一つに通知する通知手段とを有することを特徴とする計算機。

【請求項12】 複数のストレージサブシステム間においてデータを移動する方法であって、前記複数のストレージサブシステムにおいて収集された情報を前記複数のストレージシステムに接続される計算機で収集し、移動対象となるデータを前記計算機で決定し、前記決定されたデータの位置を、移動先となる前記複数のストレージサブシステムのうちの一つに通知し、

通知を受けたストレージサブシステムで前記通知されたデータの移動を行うことを特徴とするデータの移動方法。

【請求項13】 複数のストレージサブシステム間においてデータを移動する方法であって、前記複数のストレージサブシステムにおいて収集された情報を前記複数のス

トレージシステムに接続される計算機で収集し、移動対象となるデータを前記計算機で決定し、前記決定されたデータの位置を、移動先となる前記複数のストレージサブシステムのうちの一つに通知し、通知を受けたストレージサブシステムで前記通知されたデータの移動を行うことを特徴とするデータの移動方法。

【請求項14】複数のストレージサブシステム間においてデータを移動するコンピュータプログラムであって、前記複数のストレージサブシステムにおいて収集された情報を前記複数のストレージシステムに接続される計算機で収集し、

移動対象となるデータを前記計算機で決定し、前記決定されたデータの位置を、移動先となる前記複数のストレージサブシステムのうちの一つに通知するプログラムと、

通知を受けたストレージサブシステムで前記通知されたデータの移動を行うプログラムとから構成されることを特徴とするコンピュータプログラム。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明は、記憶装置に記憶されたデータを再配置する技術に関し、特に、複数の記憶装置を有する計算機システムでのデータの再配置に好適な技術に関する。

【0002】

【従来の技術】計算機システムにおける、ストレージサブシステム内に記憶されたデータを再配置する技術として、特開平9-274544号公報記載のディスクアレイシステムがある。ここで、ストレージサブシステムとは、複数の記憶装置から構成される記憶装置システムのことを言う。

【0003】ディスクアレイシステムとは、複数のディスク装置をアレイ状に配置し、各ディスク装置を並列に動作させることで、各ディスク装置に分割格納されるデータのリード/ライトを高速に行うシステムのことである。D. A. Patterson, G. Gibson, and R. H. Kats, "A Case for Redundant Arrays of Inexpensive Disks (RAID)" (in Proc. ACM SIGMOD, pp. 109-116, June 1988)に記載されているように、冗長性を付加したディスクアレイシステムには、その冗長構成に応じてレベル1からレベル5の種別が与えられている。これらの種別に加えて、冗長性無しのディスクアレイシステムをレベル0と呼ぶこともある。

【0004】ディスクアレイシステムを構成するためのコスト、ディスクアレイシステムの性能や特性等は、ディスクアレイシステムのレベルによって異なる。そこで、システムの使用目的に応じて、ディスクアレイシステムの構築の際にレベルの異なる複数のアレイ（ディスク装置の組）を混在させることも多い。このようにレベ

ルの異なるアレイを混在させたディスクアレイシステムにおいて、各レベルのアレイは、パリティグループと呼ばれる。また、ディスク装置についても性能や容量等によりコストが異なるので、最適のコストパフォーマンスを実現するシステムを構築するために、性能や容量の異なる複数種のディスク装置を用いることがある。

【0005】ディスクアレイシステムでは、データが複数のディスク装置に分散して配置されるため、ディスクアレイシステムに接続されるホストコンピュータが、論理記憶領域とディスク装置の記憶領域を示す物理記憶領域との対応付け（アドレス変換）を行っている。

【0006】特開平9-274544号公報には、ディスクアレイシステム内において、物理記憶領域間におけるデータの再配置を実行し、データの論理記憶領域に対する物理記憶領域への対応付けを、再配置前の物理記憶領域から再配置後の物理記憶領域に変更する技術が開示されている。また、ディスクアレイシステムがホストコンピュータからの各論理記憶領域に対するアクセスによる負荷状況を管理し、その実績に応じて再配置後にデータが適正配置となるように、再配置の内容を決定するという技術も開示されている。

【0007】ホストコンピュータおよびディスクアレイシステム等のストレージサブシステム間におけるデータ転送の技術としては、M. T. O' Keefe, "Shared File Systems and Fibre Channel" (in Proc. Sixth Goddard Conference on Mass Storage Systems and Technologies, pp. 1-16, March 1998)に開示された技術がある。

【0008】この技術は、高速のネットワークであるFibre Channel（以下「FC」と称する。）で複数のホストコンピュータと複数のストレージサブシステムとを接続し、FC経由でデータ共有を実現するストレージ環境、いわゆるStorage Area Network (SAN)を実現するための技術である。このように、FC経由でデータ転送を行うことにより、一般的なネットワーク経由に比べ、ホストコンピュータおよびネットワークの負荷が削減される。

【0009】高速なFCを使用しない、一般的なネットワークに接続されたストレージサブシステムに保持されているファイル等のデータを、複数のコンピュータで共有する技術としては、NFS (Network File System) が広く知られている。

【0010】NFSを用いてネットワーク間でデータ共有を行う場合には、FCを使用する場合に比べ、ファイルを共有しているコンピュータや、コンピュータとストレージサブシステムをつなぐネットワークに対する負荷が大きくなる。しかし、NFSを用いると、既存のネットワークを使用できることから、新たにFCのネットワークを敷設することと比較すると、新規設備コストを抑えられ、またファイル共有等の管理が容易である等の利点がある。

【0011】

【発明が解決しようとする課題】 上述したように、特開平9-274544号公報に開示された技術では、1つのストレージサブシステム内におけるデータの再配置が可能となる。しかしながら、複数のストレージサブシステムを有する計算機システムにおいて、異なるストレージサブシステム間でのデータの再配置を行うことはできない。また、ディスクアレイシステムはファイルを認識できないため、ファイル単位でデータの再配置を行うことができない。

【0012】 一方、SANの技術を用いれば、異なるストレージサブシステムにおける高速なデータ転送が可能となる。しかしながら、従来技術においては、SANの技術を利用してストレージサブシステム間でデータの再配置を行うことは出来なかった。

【0013】 つまり、従来技術においては、SANを構成する各ストレージサブシステムの各記憶領域の負荷状況等、データの適正な配置を決定するために必要な情報を、ホストコンピュータやシステムを使用するユーザが得ることについて、何ら考えられていない。このため、ホストコンピュータや、そのユーザが、どのようにしてデータの再配置を行えば、データの適正な配置を実現することができるのかを判断できない。

【0014】 さらに、仮にユーザが自らストレージサブシステム間におけるデータの再配置を行おうとしても、データを再配置するための未使用領域の管理等を、全てユーザが詳細に検討して行わなければならない、ユーザに対する負担は大きいものがあつた。

【0015】 また、異なるストレージサブシステム間でデータを転送すると、アプリケーションが認識するデータの位置、すなわち、アプリケーションが同じデータにアクセスするために指定すべきアクセス先が再配置の前後で変化してしまうが、この変化についても従来技術では何ら考慮されていない。

【0016】 さらに、一般的なネットワークで接続されているコンピュータ同士で、NFSを使用してデータ共有を行う場合にも、以下の問題がある。

【0017】 すなわち、従来、NFSによるデータ共有を実現するために使用されるホストコンピュータ（NFSサーバ）が、複数のストレージサブシステムを管理している場合、NFSサーバ自身は、その複数のストレージサブシステム間でのデータの物理的再配置等を行うことはできなかった。このため、NFSサーバを用いてデータ共有を行うコンピュータ毎に、共有されるデータの物理的位置を変更するといった、ストレージサブシステムの有する記憶領域についてのより細かい区別および管理を行うことができなかった。

【0018】 本発明は上記事情に鑑みてなされたものであり、本発明の目的は、NFSサーバを含めたホストコンピュータがデータの適正配置の決定に必要な情報をスト

レージサブシステムから取得できるようにし、SAN環境において、データの再配置を実現することにある。また、異なる目的としては、ユーザのストレージサブシステムの管理負担を軽減することにある。また、異なるストレージサブシステム間におけるデータの再配置を、アプリケーションが認識するデータの位置が、再配置の前後で変化しないようにして、行えるようにすることにある。さらにまた、ファイルを対象とするデータの再配置を可能とすることにある。

【0019】

【課題を解決するための手段】 前記の課題を解決するため、本発明は以下の構成とする。すなわち、計算機と、前記計算機に接続される複数のストレージサブシステムとを有し、ストレージサブシステムは、ストレージサブシステムが有する記憶領域の各々について記憶領域の使用状況を取得する取得手段を有し、計算機は、複数のストレージサブシステムから、記憶領域各々の使用状況を取得する取得手段を有すること計算機システムという構成とする。記憶領域は、論理的な記憶領域であってもよい。また、ストレージサブシステムには、計算機の指示にしたがって、データの移動を行う移動手段を付加し、計算機には、データの論理的な位置と、ストレージサブシステムにおいてデータが格納される記憶領域との対応を規定する対応テーブルと、移動手段によってストレージサブシステム間でデータが移動した場合に、対応テーブルを更新する更新手段とを付加した構成とすることもできる。

【0020】 ここで、各記憶領域の使用状況とは、例えば、その記憶領域の物理的な記憶空間の使用状況やその記憶空間へのアクセス処理に消費された処理時間等である。

【0021】 また、ストレージサブシステムで移動されるデータの単位及び計算機で管理されるデータの単位がファイル単位であることも考えられる。

【0022】 さらに、全てのストレージサブシステムが有する論理的な記憶領域全体を管理する手段と、記憶装置の特徴と論理的な記憶領域との対応関係を管理する手段とを計算機に付加した構成も考えられる。

【0023】 また、計算機に、ストレージサブシステムに格納されているデータをネットワーク間で共有する手段を付加した構成も考えられる。

【0024】

【発明の実施の形態】 図1は、本発明が適用された計算機システムの第1実施形態の構成を示す図である。

【0025】 本実施形態の計算機システムは、ホストコンピュータ（ホスト）100、ディスクアレイ200、スイッチ500、クライアント800及びローカルディスク190を有する。

【0026】 ホスト100は、ネットワークインタフェース170により、ネットワーク700を介して、クラ

クライアント800及びディスクアレイ200に接続される。ホスト100は、また、FCインタフェース160、スイッチ500及びFC600を介して、ディスクアレイ200及びローカルディスク190に接続される。

【0027】ホスト100は、ファイルシステム110、オペレーティングシステム（以下「OS」と称する。）120、マネージャ130及びアプリケーション140をホスト100自身が有する記憶領域に有する。

【0028】アプリケーションプログラム（以下、単にアプリケーションと呼ぶ）140は、OS120およびファイルシステム110を介してディスクアレイ200に対してリードやライトの要求を出す。

【0029】ホスト100及びクライアント800としては、一般的な電子計算機が用いられる。ファイルシステム110等のプログラムは、ホスト100の外部にあるローカルディスク190に記憶され、必要に応じてホスト100に読み込まれて実行される。

【0030】ホスト100がその内部に記憶装置を有する場合には、当該記憶装置をローカルディスク190として使用することも考えられる。

【0031】ローカルディスク190には、OS120及びファイルシステム110が使用する論理ユニット（以下、「LU」と称する）論理位置名テーブル191及びLU内アドレス論理位置名テーブル195等の各種管理情報が格納されている。LU論理位置名テーブル191は、アプリケーション140がディスクアレイシステム200のデータにアクセスするときに指定する論理位置名と、論理位置名により特定されるデータを格納するLUとの対応を示す情報を保持したテーブルである。

【0032】LU内アドレス論理位置名テーブル195は、アプリケーション140がディスクアレイシステム200のデータにアクセスするときに指定する論理位置名と、論理位置名により特定されるデータのLU内アドレスとの対応を示す情報を保持したテーブルである。

【0033】ディスクアレイ200は、制御部300、複数のディスク装置210、ネットワークインタフェース270及びFCインタフェース260を有する。

【0034】制御部300は、処理を実行するためのCPU310、メモリ320及びキャッシュメモリ330を有する。

【0035】メモリ320には、論理/物理対応情報321、論理ボリューム使用状況322及びコピー領域管理テーブル323が格納される。これらの情報の詳細については後述する。

【0036】本実施形態では、 n 台（ n は2以上の整数）のディスク装置210でアレイ（以下「RAID」と称する。）が構成されており、この n 台のディスク装置210によるRAIDをパリティグループ220と呼ぶ。

【0037】RAIDとしては、1つのパリティグループ220に含まれる n 台のディスク装置210のうち、 $n-1$

台のディスク装置210の格納内容から生成される冗長データ（以下「パリティ」と称する。）を残りの1台に格納する構成や、 $n/2$ 台に格納されている内容を残りの $n/2$ 台がコピーしたミラーディスク（RAIDレベル1）構成が考えられる。また、各パリティグループ220を、ホスト100からのアクセス対象の1単位とみなすことができる。

【0038】本実施形態においては、ディスクアレイ200を構築する各パリティグループ220の性能、信頼性、特性などの属性は、同一であってもよいし、あるいは、異なっていてかまわない。

【0039】制御部300は、ホスト100がリード/ライトする論理ボリュームとディスク装置210の記憶領域を示す物理アドレスとの対応付け（アドレス変換）を行い、ホスト100に論理ボリュームを提供する。ディスクアレイ200は、アドレス変換において、複数の論理ボリュームを結合して1つのLUとしてホスト100に提供することもできる。すなわち、ディスクアレイ200は、少なくとも1つの論理ボリュームからなるLUをホスト100に提供する。ホスト100は、LUに対してリード/ライトを行う。

【0040】本実施形態では、複数のディスクアレイ200間における、ディスクアレイ200の使用状況を考慮したデータの物理的再配置を可能とする。具体的には、ディスクアレイ200は、ホスト100からのリード/ライトに対するディスク装置210の使用状況を取得する。ホスト100は、複数のディスクアレイ200各々が取得した使用状況を収集し、ユーザに提示する。さらに、ホスト100は、ディスクアレイ200の使用状況の提示を受けたユーザからの指示等に応じ、ローカルディスク190内のLU論理位置名テーブル191を変更すると共に、ディスクアレイ200がLUに格納しているデータをコピーする。これにより、複数のディスクアレイ200間におけるLUの再配置が行われる。そして、ディスクアレイ200の使用状況を考慮したデータの再配置を可能とすることにより、データの適正配置が可能となる。

【0041】図2は、ディスクアレイ200が、ホスト100からのリード/ライト要求に回答して行うリード/ライト処理、及びディスクアレイ200が、ディスク装置210の使用状況を取得する使用状況取得の処理の手順を示すフロー図である。使用状況取得の処理は、随時、又はホスト100からのリード/ライト要求時に行われる。

【0042】ホスト100のアプリケーション140は、ファイル論理位置名によってファイルを指定し、ファイルに対するリード/ライトをOS120に要求する。OS120は、ファイルシステム110にファイルのリード/ライトを要求する。

【0043】ファイルシステム110は、FCインターフ

エース160を介してローカルディスク190にアクセスし、指定されたファイルが格納されているLU番号をLU論理位置名テーブル191から求める。ファイルシステム110は、指定されたファイルが格納されているLU内アドレス等を、LU内アドレス論理位置名テーブル195から求める。

【0044】ホスト100は、FCインターフェース160を介して、求めたLU番号のLUを提供するディスクアレイ200に対し、LU番号やLU内アドレスを伴うSmall Computer System Interface (SCSI) 規格のリードコマンド、あるいはライトコマンドを発行する。

【0045】アプリケーション140が、論理ドライブ名、ディレクトリ名及びファイル名によるファイルの論理位置までのパスの記述によりファイルを指定するシステムでは、論理位置（論理ドライブやディレクトリやファイル）へのパスの記述が、ファイルの論理位置の論理位置名となる。一般的には、論理位置名とは、アプリケーションがアクセス対象の指定に使用する論理位置の情報である。

【0046】ファイルシステム110は、各論理位置を管理するため、ディレクトリ構造などの各論理位置間の階層的な論理構造を管理する他、各論理位置の論理位置名とLU番号との対応をLU論理位置名テーブル191に記述し管理する。また、各論理位置の論理位置名とLU内アドレスとの対応をLU内アドレス論理位置名テーブル195に記述し管理する。なお、LU番号は、そのLU番号のLUを提供するディスクアレイ200も表す（ステップ1000）。

【0047】ホスト100からリード/ライトのコマンドを受領した制御部300は、メモリ320内の論理/物理対応情報321を用いて、コマンドで指定されているLU番号により特定されるLUを構成する論理ボリュームを特定する。制御部300は、論理ボリュームに対応するディスク装置210の領域を求め、コマンドで指定されているLU内アドレスの物理アドレスへの変換を行う。

【0048】論理/物理対応情報321は、図3に示すように、LUとディスク装置210の物理アドレスとの対応関係についての情報を保持するテーブルである。

【0049】図中、LU番号5001およびLU内アドレス5002は、ホスト100のファイルシステム110がリード/ライト処理で指定するLU番号及びLU内アドレスを示す。論理ボリューム番号5003は、LU番号5001で特定されるLUに対応する論理ボリュームの番号である。論理ボリュームアドレス5004は、LU内アドレス5002に対応する論理ボリューム内のアドレスである。

【0050】物理アドレスは、データとパリティが格納されるディスク装置210上の領域を示すアドレスである。物理アドレスは、パリティグループ番号5005、データおよびパリティ各々に対するディスク装置番号5

006及びディスク装置内アドレス5007を有する。パリティグループ番号5005は、個々のパリティグループ220を示す。ディスク装置番号5006は、個々のディスク装置210を示す。ディスク装置内アドレス5007は、ディスク装置210内での領域を示すアドレスである（ステップ1010）。

【0051】データのリードの場合、制御部300は、アドレス変換で得た物理アドレスに基づいて、ディスク装置210のデータを読み出し、ホスト100に転送する。データのライトの場合、制御部300は、ホスト100から転送されたデータ及びデータに関連して生成したパリティを、アドレス変換で得たディスク装置210の物理アドレスの位置に格納する（ステップ1020）。

【0052】、リード/ライト処理を終了した制御部300は、使用状況取得処理を実行する。この処理では、制御部300は、リード/ライト処理でのリード/ライト種別やシーケンシャル/ランダムアクセス種別を判別し、メモリ320のリード/ライト対象となった論理ボリュームの論理ボリューム使用状況322を更新する。論理ボリューム使用状況322は、ディスクアレイ200に含まれるLUの使用状況についての情報を保持したテーブルである。論理ボリューム使用状況322の一例を、図4に示す。

【0053】論理ボリューム使用状況322には、論理ボリューム毎に、論理ボリューム番号5101及びリード/ライト種別およびシーケンシャル/ランダムアクセス種別毎のディスク使用時間（マイクロ秒単位）5102が記述される。ここでは、リード/ライト対象となった論理ボリュームの論理ボリューム番号5101に対応する、ディスク使用時間5102に、リード/ライトに要した時間が加算される（ステップ1030）。

【0054】図5は、ホスト100が、各ディスクアレイ200からディスク装置210の使用状況を収集する使用状況収集処理の手順を示すフロー図である。この処理は、随時行われる。

【0055】ホスト100のマネージャ130は、FCインターフェース160を介し、コマンドボリュームに対して、情報収集用のパラメータをライトデータとするSCSI規格のライトコマンドを発行する。コマンドボリュームは、ディスクアレイ200が有する情報転送用のLUであって、対応する物理領域が指定されない論理ボリュームである。（ステップ1100）。

【0056】制御部300は、発行されたコマンドがコマンドボリュームに対するライトコマンドであることを確認すると、ホスト100から転送された情報収集用のパラメータに含まれるオペレーションコードから、要求された情報を判別する。制御部300は、要求された情報をメモリ320上に用意する（ステップ1110）。制御部300は、FCインターフェース260を介して、

ホスト100にライトの完了を報告する(ステップ1120)。

【0057】完了報告を受けたホスト100のマネージャ130は、FCインターフェース160を介して、ディスクアレイ200のコマンドボリュームに、SCSI規格のリードコマンドを発行する(ステップ1130)。

【0058】制御部300は、コマンドボリュームに対するリードコマンドを受領すると、メモリ320上に用意した情報を、FCインターフェース260を介してホスト100に転送する(ステップ1140)。制御部300は、FCインターフェース260を介してホスト100にリードの完了を報告する(ステップ1150)。

【0059】ステップ1100でライトされる情報収集用のパラメータ及びステップ1110で用意される情報には、論理ボリューム情報、パリティグループ情報及び使用状況情報の3種類の情報が含まれる。

【0060】ステップ1100でライトされる情報収集用のパラメータが、図6に示すような論理ボリューム情報のパラメータである場合、制御部300は、その0～1バイト目で指定された論理ボリューム番号5201で特定される論理ボリュームについて、図7に示すような論理ボリューム情報(ディスクアレイ200内のその論理ボリュームの構成を示す情報)を用意する。

【0061】図7に示す論理ボリューム情報において、8～47バイト目には、その0～1バイト目に記述されている論理ボリューム番号5201で特定される論理ボリュームの各種情報5202が記述される。49～121バイト目には、その論理ボリュームが属するLUを構成する各論理ボリュームの情報5203が記述される。

【0062】情報収集用のパラメータが、パリティグループ情報のパラメータの場合、制御部300は、パラメータで指定された論理ボリュームが属するパリティグループ220のパリティグループ情報(RAIDの構成、ディスク装置210の型名等、ディスクアレイ200内のそのパリティグループ220の構成を示す情報)を用意する。

【0063】情報収集用のパラメータが、ディスク装置210の使用状況を確認するためのパラメータの場合、制御部300は、パラメータで指定された論理ボリュームの使用状況情報(ディスクアレイ200内のリソースの使用状況、例えば論理ボリュームが占有される時間、論理ボリュームの各種コマンド受領回数やキャッシュメモリ330のヒット回数等の情報、プロセッサ310の占有時間及び内部バスの占有時間等の情報等)を用意する。

【0064】制御部300は、あらかじめ、論理ボリューム毎に、各種コマンド受領回数やキャッシュメモリ330のヒット回数やプロセッサ310の占有時間や内部バスの占有時間等を取得している。マネージャ130は、例えば複数回取得した占有時間の平均を取得間隔で

割ることにより、単位時間あたりの占有時間率を求めることができる。

【0065】制御部300は、論理ボリューム情報やパリティグループ情報を生成する際に、論理/物理対応情報321の一部あるいは全部を使用する。マネージャ130は、各型のディスク装置210の性能に関する情報を保持しており、ディスク装置210の型名を基に、パリティグループ220を構成するディスク装置210の性能を得ることができる。

【0066】また、ホスト100のマネージャ130は、LUに対しSCSI規格のINQUIRYコマンドを発行して応答データを得ることで、この応答データからLUに属する論理ボリューム番号を得ることもできる。

【0067】図8は、ホスト100が再配置すべきデータを決定する再配置対象決定処理の手順を示すフロー図である。本処理は、ユーザが再配置すべきデータを検索する際に使用するアプリケーションが実行された時に実行される。

【0068】ホスト100のマネージャ130は、OS120が使用しているLU及び使用していないLU(空きLU)を、例えばローカルディスク190に格納されているLU論理位置名テーブル191から判定する。マネージャ130は、OS120が使用している各LUについて、LUが属するディスクアレイ200における各論理ボリュームの使用状況や、LUに対応する論理ボリュームの使用状況等を計算する。この計算には、INQUIRYコマンドを発行して得られるLUに属する論理ボリューム番号、使用状況収集処理で得られる各ディスクアレイ200における論理ボリューム情報、パリティグループ情報および論理ボリュームの使用状況等が使用される(ステップ1200)。

【0069】マネージャ130は、使用状況等の計算結果を、各論理ボリュームが属するパリティグループ220の属性(RAID構成、ディスク装置210の型名又はディスク装置210の性能等)等と共にユーザに提示する(ステップ1210)。

【0070】マネージャ130は、各LUについて、INQUIRYコマンドを発行して得られた各LUに属する論理ボリューム番号、使用状況収集処理で得られた各ディスクアレイ200における論理ボリューム情報、パリティグループ情報及び論理ボリュームの使用状況等とから、各空きLUが対応する各論理ボリュームの使用状況等を計算する(ステップ1220)。この計算結果が、各空きLUに関連するパリティグループ220の属性等と共にユーザに分類されて提示される(ステップ1230)。

【0071】使用状況等の情報は、ホスト100あるいはホスト100にネットワーク接続された他の計算機で表示することもできる。

【0072】ユーザは、各ディスクアレイ200の各LUについての情報を参照し、データを再配置すべきLU(再

配置元LU)及びデータの再配置先のLUを決定する。ただし、ユーザではなく、ホスト100のマネージャ130が、各LUについての情報から自動的にデータ再配置元又は再配置先を決定してもよい。再配置の決定は、たとえば、再配置後に、ディスクアレイ200間での負荷分散、パリティグループ220間での負荷分散、高性能を要求するファイルが存在するLUの高性能パリティグループ220への配置等が実現されるように行なわれる。再配置先LUのサイズは、再配置元LUのサイズ以上でなければならない。各LUのサイズは、SCSI規格のREAD CAPACITYコマンドで取得することができる(ステップ1240)。

【0073】図9は、ホスト100が行う、データの再配置処理の手順を示すフロー図である。ホスト100は、再配置を決定したユーザの指示、例えば再配置を指示する実行コマンドの入力等があった場合に本処理を実行する。

【0074】ユーザからの指示が入力されたホスト100のマネージャ130は、ファイルシステム110に再配置元LUのロックを指示する(ステップ1300)。ファイルシステム110は、ロック指示に応じて、再配置元LUへのリード/ライト要求の受付を禁止する(ステップ1310)。

【0075】次に、マネージャ130は、ファイルシステム110に、再配置元LUについてのキャッシュメモリのフラッシュを指示する(ステップ1320)。ファイルシステム110は、再配置元LUに格納されるデータであって、ホスト100上のメモリにキャッシュされていて且つディスクアレイ200に未だライトされていないデータを、ディスクアレイ200の再配置元LUにライトする(ステップ1330)。

【0076】マネージャ130は、ファイルシステム110に、再配置元LUについてのキャッシュの無効化を指示する(ステップ1340)。ファイルシステム110は、再配置元LUに格納されるデータであってホスト100上のメモリにキャッシュされているデータを無効にする(ステップ1350)。

【0077】LUのロック、キャッシュメモリのフラッシュ及び無効化の処理は、LUのアンマウントの処理に相当する。

【0078】マネージャ130は、再配置先LUが存在するディスクアレイ200に、再配置元LUから再配置先LUへのデータのコピーを指示する。この指示は、使用状況収集処理と同様、再配置先LUが存在するディスクアレイ200のコマンドボリュームに、コピー指示オペレーションコードや再配置元LUや再配置先LU等のコピー指示のパラメータを含んだライトコマンドを発行することで行われる(ステップ1360)。ディスクアレイ200は、後述のコピー処理を開始し、コピー指示の受領をマネージャ130に通知する(ステップ1370)。

【0079】マネージャ130は、ローカルディスク190に格納されているファイルシステム110が使用するLU論理位置名テーブル191を書き換え、再配置元LUと再配置先LUとの論理位置名を入れ替える。入れ替えられるLU論理位置名テーブル191の例を、図10及び図11に示す。

【0080】図中、ディスクアレイ番号、IDおよびLUNは、LU番号6001を特定するために必要な情報である。図10は、論理位置名をディレクトリ形式で示したものであり、図11は、論理位置名をドライブ形式で示したものである。いずれも、アプリケーション140が使用する記憶領域としてのLUの論理位置を示している(ステップ1380)。マネージャ130は、ファイルシステム110に、LU論理位置名テーブル191の更新(再読み込み)及びステップ1300で指示したロックの解除を指示する(ステップ1390)。

【0081】ファイルシステム110は、LU論理位置名テーブル191を再度読み込んで情報を更新する(ステップ1400)。ファイルシステム110は、ロックを解除してリード/ライト要求の受け付けを再開する(ステップ1410)。

【0082】ステップ1400及び1401の処理は、LUのマウント処理に相当する。

【0083】ステップ1401の処理が実行された後は、ファイルシステム110のリード/ライトの対象となるLUが再配置の対象であるLUであれば、ファイルシステム110のリード/ライト処理は、ステップ1380において情報が入れ替えられた再配置先LUに対して行われる。

【0084】図12は、再配置処理において、ディスクアレイ200が、ホスト100からコピー指示を受けた際に行うコピー処理の手順を示すフロー図である。

【0085】再配置先LUが存在するディスクアレイ200が、FCインターフェース260を介してホスト100からコピー指示を受け取ると、制御部300は、コピー指示で指定された再配置先LUについてのコピー領域管理テーブル323をメモリ320上に用意する。

【0086】図13は、コピー領域管理テーブル323の内容を示す図である。コピー領域管理テーブル323は、コピーされるデータの範囲、大きさ等の情報が登録されているテーブルである。

【0087】図中、コピー先LU番号6101及びコピー元LU番号6102は、FC600のネットワーク内において再配置先LUと再配置元LUを一義的に示す番号を格納する領域である。具体的には、ホスト100からコピー指示のパラメータとして指定された8バイトの番号(WORLD WIDE NAME)、3バイトの番号(N_PORT ID)、SCSI規格のターゲットIDもしくはLUNが格納される。コピーブロック数6103には、コピーする領域のブロック(最小リード/ライト単位)の数であり、コピー領域の大き

さを示すデータが格納される。ビットマップ 6104 のビットには、LU のコピー対象領域の各ブロックが割り当てられる。ビットが「1」である場合は未コピーを示し、「0」である場合はコピー済を示す。初期時は、コピー対象領域に対応するすべてのビットが 1 に設定される (ステップ 1500)。

【0088】制御部 300 は、コピー指示の受領をホスト 100 に通知する。この通知は、コピー指示を実際に受領してから、コピー領域管理テーブル 323 の設定後、実際にコピーを行う前の時点で行われる。このため、コピー指示の受領から当該通知までの時間は短い (ステップ 1510)。

【0089】制御部 300 は、FC インターフェース 260 を介して再配置元 LU から格納すべきデータをリードし、再配置先 LU に格納するコピーを行う (ステップ 1520)。

【0090】制御部 300 は、LU のコピー対象領域について、コピー済の領域に対応するビットマップ 6104 のビットを順次 0 に変更する (ステップ 1530)。制御部 300 は、対象となる LU 全体のコピーが終了したら、コピー処理を終了する (ステップ 1540)。

【0091】再配置元 LU が存在するディスクアレイ 200 と再配置先 LU が存在するディスクアレイ 200 とが同一の場合には、ディスクアレイ 200 内で LU のコピーが行われる。

【0092】ホスト 100 からの再配置対象 LU へのリード/ライトは、再配置対象 LU のデータがコピー中であっても、再配置先 LU、すなわち再配置先 LU の存在するディスクアレイ 200 に対して行われる。

【0093】図 14 は、再配置先 LU の存在するディスクアレイ 200 が、データの再配置におけるコピー処理の最中に、再配置の対象となる LU に対するリード/ライトコマンドを受けた場合における処理の手順について示すフロー図である。

【0094】ディスクアレイ 200 が、FC インターフェース 260 を介してリードコマンドを受け取ると、制御部 300 は、リード対象範囲とテーブル 323 のビットマップ 6104 とを比較する (ステップ 1610)。リード対象領域に未コピーの領域が含まれている場合には、制御部 300 は、リード対象領域のデータを優先して読み出してコピーを済ませる (ステップ 1630)。制御部 300 は、ビットマップ 6104 のリード対象領域に対応するビットをコピー済みに更新する (ステップ 1640)。制御部 300 は、ディスクアレイ 200 内のコピーしたデータをホスト 100 に転送する (ステップ 1650)。リード対象領域がすべてコピー済であれば、制御部 300 は、ディスクアレイ 200 内のコピー済みのデータをホスト 100 に転送する (ステップ 1650)。

【0095】制御部 300 は、FC インターフェース 260

0 を介してライトコマンドを受け取ると、ホスト 100 から転送されたデータについて、ライト対象領域にライトを行う (ステップ 1670)。制御部 300 は、コピー領域管理テーブル 323 のビットマップ 6104 のライト対象領域に対応するビットをコピー済みに更新する (ステップ 1680)。制御部 300 は、残りの未コピー領域のコピーを継続する (ステップ 1690)。

【0096】以上の処理により、再配置先 LU が存在するディスクアレイ 200 は、データ再配置におけるコピー処理中であっても、ホスト 100 からのリード/ライトを処理することができる。

【0097】なお、このリード/ライトの処理の際、制御部 300 は、同時に、先に説明した使用状況取得処理も行う。

【0098】また、ホスト 100 のマネージャ 130 は、データ再配置におけるコピー処理中に、ディスクアレイ 200 のコマンドボリュームにコピー進捗取得のためのパラメータを含むデータのライトコマンドを発行し、ディスクアレイ 200 が用意したデータをリードすることで、コピーの進捗情報等をディスクアレイ 200 に問い合わせることができる。

【0099】具体的には、コマンドボリュームに対するライトコマンドを受け付けた制御部 300 は、コマンドボリュームにライトされたパラメータを確認する。制御部 300 は、コピー領域管理テーブル 323 を参照してパラメータに対応するコピーの進捗率などの情報をメモリ 320 上に用意し、ライト完了をホスト 100 に通知する。マネージャ 130 は、コマンドボリュームに対するリードを行う。制御部 300 は、ホスト 100 のリードに対して、メモリ 320 上に用意したデータを転送することによって、コピーの進捗等の問い合わせに答える。

【0100】本実施形態によれば、複数のディスクアレイ 200 間における LU の再配置によるデータの適正配置を、アプリケーション 140 にとって再配置前後で論理的に等価となるように、すなわち、アクセス対象のアクセスにアプリケーションが使用すべき論理位置名が変化しないようにしつつ実現できる。

【0101】また、本態様によれば、計算機は、ディスクアレイから取得した、各記憶領域の物理的な記憶装置資源の使用状況を、例えば記憶装置資源の負荷分散等の観点による、データの適正配置の決定に用いることができる。したがって、この情報を用いて、例えば異なるストレージサブシステム間でデータを再配置することにより、データの適正配置を行うことができる。

【0102】なお、本実施形態では、複数のディスクアレイ 200 間におけるデータの再配置について説明した。しかし、再配置対象データを格納するストレージサブシステムは、ディスクアレイサブシステムでなくともよい。磁気ディスク装置、光磁気ディスク装置、磁気テ

ープ装置又は半導体ディスク装置などを用いた他の種類のストレージサブシステムであってもよい。

【0103】尚、マネージャ130は、FC600経由ではなく、ネットワーク700経由で、例えばSimple Network Management Protocol (SNMP) で規定されているプロトコルを用いて情報の収集や指示を行ってもよい。

【0104】本実施形態では、ディスクアレイ200の制御部300が取得する論理ボリューム使用状況322が使用時間の累積値である場合について説明した。しかし、制御部300が単位時間毎の使用時間を使用率の形式にしてメモリ320に蓄積し、これを論理ボリューム使用状況322として、ホスト100のマネージャ130が収集するようにしてもよい。

【0105】図15は、本発明が適用された計算機システムの第2実施形態の構成を示す図である。

【0106】図示するように、本実施形態の計算機システムは、ローカルディスク190にLU領域範囲テーブル192を格納し、スイッチ500にコピー制御部510を設けた構成を有している点が、第1実施形態の計算機システムと異なる。

【0107】本実施形態では、ディスクアレイ200がディスク装置210の使用状況を取得し、ホスト100が複数のディスクアレイ200から使用状況を収集し、使用状況を計算機システムのファイルに基づく分析も含めてユーザに提示する。ホスト100は、ファイル管理のためのデータ（以下「メタデータ」と称する）を変更する。スイッチ500は、ホスト100の指示に基づいて、ディスクアレイ200に格納されているデータをコピーする。これにより、複数のディスクアレイ200間におけるファイルの再配置を可能とし、データの適正配置を行えるようにする。

【0108】第1実施形態においては、ホスト100のファイルシステム110は、各LUを、使用中のものと使用していないものとに区別して管理した。本実施形態では、ファイルシステム110は、全てのLUを使用し、全てのLUの領域の集合を単一領域（以下、「統合領域」と称する。）として管理する。また、統合領域上のファイルを、後述するメタデータで管理する。メタデータは、統合領域の既定の位置に格納される。

【0109】図16は、本実施形態において、ファイルシステム110が統合領域を管理するために用いるLU領域範囲テーブル192の例を示した図である。LU領域範囲テーブル192は、統合領域の範囲と各LU内領域の範囲との対応を示す情報を保持している。

【0110】図中、領域内アドレス6301には、統合領域内でのアドレスが格納される。LU番号6302は、ディスクアレイ番号、ID及びLUNを含み、領域内アドレス6301に格納されるLUを示す。LU内アドレス6303は、対応するLU番号6302で特定されるLU内でのアドレスが格納される図17は、ホスト100がリード/

ライトを行う場合の処理の手順を示すフロー図である。

【0111】前提として、ホスト100のアプリケーション140は、ファイルシステム110が管理するファイルの論理位置を指定して、ディスクアレイ200が格納するデータにリードやライトを行うものとする。また、ファイルシステム110は、データをファイルとして管理するために、メタデータをディスクアレイ200に格納している。

【0112】なお、メタデータはディスクアレイ200に格納されているが、ファイルシステム110の管理に基づき、ホスト100上のメモリにキャッシュされている場合もある。以下、メタデータがホスト100上のメモリにキャッシュされている場合で説明する。

【0113】図18は、メタデータの内容を示す図である。

【0114】図示するように、メタデータには、各ファイルの作成日時、更新日時、アクセス日時、属性、ファイル論理位置名、セキュリティ情報、及びファイル位置等が含まれる。各ファイルに対応する統合領域内の範囲は、ファイル位置6407に格納された情報で示される。

【0115】ホスト100のアプリケーション140は、ファイル論理位置名によってファイルを指定し、ファイルに対するリード/ライトをOS120に要求する（ステップ1700）。OS120は、ファイルシステム110に、ファイルのリード/ライトを要求する（ステップ1710）。ファイルシステム110は、キャッシュメモリされているメタデータを参照し、メタデータ及びLU領域範囲テーブル192の情報から、指定されたファイルの位置（LUおよびLU内アドレス）を得る（ステップ1720）。

【0116】要求がライト要求である場合、ファイルシステム110は、さらにメタデータの更新を行う（ステップ1740）。ファイルシステム110は、ステップ1720で得たファイルの位置が示す領域内のリード/ライトをディスクアレイ200に対して行い（ステップ1750）、キャッシュされたメタデータ及びディスクアレイ200のメタデータを更新する（ステップ1760）。

【0117】ステップ1740及び1760でのメタデータの更新は、アクセスされたファイルについて、作成日時6401、更新日時6402、アクセス日時6403、属性6404、ファイル論理位置名6405、セキュリティ情報64060、及びファイル位置6407等に格納された情報を、アクセス内容に応じて更新することで行われる。例えば、ライトによりファイルサイズが増減する場合は、これに合わせて、メタデータのファイル位置6407が示す領域内の範囲が増減される。また、ファイルが新規に作成される場合は、メタデータに新規ファイルのエントリが追加され、ファイルが削除さ

れる場合は対応するエントリが削除される。

【0118】本実施形態において、制御部300は、第1実施形態と同様の使用状況取得処理を行う。また、ホスト100のマネージャ130は、第1実施形態と同様の使用状況収集処理を行う。

【0119】図19は、ホスト100が行うファイル単位の再配置対象決定処理の手順を示すフロー図である。

【0120】ホスト100のマネージャ130は、統合領域に存在する各ファイルについて、ファイルとLUとの対応を、ファイルシステム110に問い合わせる（ステップ1800）。ファイルシステム110は、キャッシュされたメタデータ及びLU領域範囲テーブル192を用いて、問い合わせに答える（ステップ1810）。

【0121】マネージャ130は、ディスクアレイ200毎の各論理ボリュームの使用状況、各LUの各論理ボリュームの使用状況及びファイル毎の各論理ボリュームの使用状況等を計算する。この計算には、INQUIRYコマンドによって得られた各ディスクアレイ200における各LUに属する論理ボリューム番号、使用状況収集処理で得られた各ディスクアレイ200における論理ボリューム情報及びパリティグループ情報及び論理ボリュームの使用状況等が使用される（ステップ1820）。マネージャ130は、計算結果を、各論理ボリュームが属するパリティグループ220の属性等と共にユーザに提示する。すなわち、ホスト100は、使用状況に関する情報を、ディスクアレイ200、論理ボリューム、LU、ファイルといった各種の視点でユーザに提供する（ステップ1830）。

【0122】マネージャ130は、各ディスクアレイ200が提供するLUや論理ボリュームについて利用可能な空き領域を計算し、ユーザに提示する（ステップ1840）。マネージャ130は、各ディスクアレイ200が提供するLUや論理ボリュームについて利用可能な空き領域を、ファイルシステム110に問い合わせる（ステップ1850）。ファイルシステム110は、キャッシュされたメタデータ及びLU領域範囲テーブル192を参照して、ファイルが存在しない空き領域を特定し、マネージャ130に答える（ステップ1860）。マネージャ130は、使用状況収集処理で得た各種使用状況等から、空き領域の論理ボリュームの使用状況等を、論理ボリュームやパリティグループ220の属性等と共にユーザに分類して提示する（ステップ1870）。

【0123】使用状況や空き領域の情報は、ホスト100またはホスト100にネットワークで接続された他の計算機で表示することができる。ユーザは、これらの情報より再配置すべきファイルと再配置先の空き領域とを決定する。マネージャ130は、これらの情報から、自動的に同様の再配置対象や空き領域を決定してもよい（ステップ1880）。

【0124】ホスト100のファイルシステム110

が、OS120やアプリケーション140からの各ファイルへのリード/ライト要求頻度（アクセス頻度）を監視して統計情報を生成し、ステップ1830でユーザに提示するようにしてもよい。

【0125】これにより、ユーザは、ホスト100での各ファイルのアクセス頻度を勘案して再配置すべきファイルを決定することができる。

【0126】図20は、ホスト100が、再配置対象決定処理の結果を受けて行う再配置処理の手順を示すフロー図である。本処理は、基本的には、図9に示すLU単位の再配置決定処理の手順において、LUをファイルに、ディスクアレイ200をスイッチ500に読み替えた処理と同じである。以下、図9とは異なる部分についての説明する。

【0127】マネージャ130は、ファイルシステム110に、再配置先の空き領域についての領域の使用予約を指示する（ステップ1940）。ファイルシステム110は、指定された再配置先領域が確保されるよう、キャッシュされたメタデータを更新する（ステップ1950）。マネージャ130は、ファイルシステム110に、メタデータのキャッシュメモリのフラッシュを指示する（ステップ1960）。ファイルシステム110は、ホスト100上のメモリにキャッシュメモリしてあるメタデータを、ディスクアレイ200にライトする（ステップ1970）。

【0128】マネージャ130は、メタデータを書き換え、指定されたファイルの位置を、再配置元領域から再配置先領域へ入れ替える。これにより、再配置元の領域を空き領域とする（ステップ2010）。マネージャ130は、ファイルシステム110に、メタデータについて、キャッシュの無効化を指示する（ステップ2020）。ファイルシステム110は、ホスト100上のメモリにキャッシュしてあるメタデータを無効にする（ステップ2030）。

【0129】以降、ファイルシステム110が、ファイルにリード/ライトする場合には、再配置先領域にコピーされたデータに対して正常にリード/ライトを行うことができる。

【0130】本実施形態によれば、複数のディスクアレイ200間でのファイルの適正配置を、アプリケーション140にとって再配置前後で論理的に等価となるように行うことが可能となる。

【0131】図21は、本発明が適用された計算機システムの第3実施形態の構成を示す図である。

【0132】本実施形態の計算機システムは、クライアント800が、FCインタフェース860及びネットワークインタフェース870を有する。そして、クライアント800がFCインターフェース860を介してFC600経由でホスト100、ディスクアレイ200及びスイッチ500に接続され、かつネットワークインターフェー

ス 870 を介してネットワーク 700 経由でホスト 100 およびディスクアレイ 200 に接続される点が、第 2 実施形態の計算機システムと異なる。本実施形態では、複数のクライアント 800 とホスト 100 とが、ディスクアレイ 200 上のファイルを共有する。クライアント 800 は、OS 820 とアプリケーション 840 を有する。クライアント 800 は一般的な電子計算機である。

【0133】第 2 実施形態と同様に、本実施形態のファイルシステム 110 は、全ての LU を使用し、全ての LU の領域を集合して単一の統合領域として管理する。そして、統合領域上のファイルを、第 2 実施形態と同様にメタデータにより管理する。

【0134】クライアント 800 が、ディスクアレイ 200 に格納されているファイルへアクセスする処理について説明する。

【0135】図 22 は、クライアント 800 がディスクアレイ 200 に格納されているファイルのリードを行う場合の処理の手順を示すフロー図である。

【0136】クライアント 800 のアプリケーション 840 は、OS 820 にファイルのリードを要求する（ステップ 2100）。OS 820 は、ネットワークインターフェース 870 あるいは FC インターフェース 860 を介して、ホスト 100 のファイルシステム 110 にファイルのリードを通知する（ステップ 2110）。

【0137】ファイルのリードの通知を受けたファイルシステム 110 は、ファイルが格納されている LU および LU 内アドレスを、メタデータと LU 領域範囲テーブル 192 とを参照して求める（ステップ 2120）。ファイルシステム 110 は、ファイルが格納されている LU の LU 内アドレスを他のクライアント 800 からのライトに対してロックする（ステップ 2130）。ファイルシステム 110 は、ホスト 100 のキャッシュメモリにあるメタデータをフラッシュする（ステップ 2140）。ファイルシステム 110 は、クライアント 800 の OS 820 に、ファイルが格納されている LU および LU 内アドレスと、メタデータの格納されている LU および LU 内アドレスとを返答する（ステップ 2150）。

【0138】返答を受けたクライアント 800 の OS 820 は、リードの対象となるファイルが格納されている LU が存在するディスクアレイ 200 に対し、FC インターフェース 860 を介して、ファイルが格納されている LU 内アドレスに対するリードを行って、アプリケーション 840 からの要求を処理する（ステップ 2160）。

【0139】クライアント 800 から要求されたデータのリード処理が終了したら、OS 820 は、ホスト 100 のファイルシステム 110 から通知された LU および LU 内アドレスにあるメタデータ上のファイルのアクセス日時を更新する（ステップ 2170）。OS 820 は、ファイルシステム 110 に、ネットワークインターフェース 870 または FC インターフェース 860 を介して、処理の

完了を通知する（ステップ 2180）。

【0140】完了通知を受けたファイルシステム 110 は、ホスト 100 上のメタデータのキャッシュメモリを無効化し（ステップ 2190）、ステップ 2130 で行ったロックを解除する（ステップ 2200）。

【0141】図 23 は、ライトを行う場合の処理の手順を示すフロー図である。

【0142】ライト処理は、図 22 のリード処理において、リードをライトに置き換えた処理とほぼ同一である。以下、異なる部分について説明する。

【0143】ファイルシステム 110 は、ライトで増加する可能性のあるファイル使用領域のための領域の予約をメタデータに記述する（ステップ 2340）。ファイルシステム 110 は、クライアント 800 の OS 820 に、ファイルが格納されている LU および LU 内アドレス（ライトで増加する可能性のあるファイル使用領域のために予約した領域を含める）と、メタデータが格納されている LU 及び LU 内アドレスとを返答する。なお、ライトで増加する可能性のあるファイル使用領域の増加量は、クライアント 800 の OS 820 からのライトの通知に含まれているものとする（ステップ 2360）。

【0144】返答を受けた OS 820 は、ライトの対象となるファイルが格納されている LU が存在するディスクアレイ 200 に対し、FC インターフェース 860 を介して、ファイルが格納されている LU 内アドレスに対するライトを行い、アプリケーション 840 からの要求を処理する（ステップ 2370）。

【0145】このようにして、クライアント 800 のアクセスを処理することにより、クライアント 800 およびホスト 100 は、ディスクアレイ 200 に格納されているファイルを矛盾なく共有して使用することができ、なお、ホスト 100 自身のファイルアクセスも、クライアント 800 によるファイルアクセスと同様に処理される。

【0146】次に、本実施形態でのファイルの再配置について説明する。

【0147】本実施形態でのファイルの再配置に関する処理（使用状況取得処理、使用状況収集処理、再配置対象決定処理および再配置処理）は、第 2 実施形態と同様である。ただし、アプリケーション 840 が要求するデータのリード/ライト処理でファイルがロックされている間、再配置処理は実行されない。また、図 20 に示す再配置処理のステップ 1920 及び 1930 におけるファイルのキャッシュメモリのフラッシュと、ディスクアレイ 200 への書き戻しは、ファイルシステム 110 がそのファイルをキャッシュメモリしているクライアント 800 に対して指示し、これを行わせる。

【0148】本実施形態によれば、ディスクアレイ 200 に格納されているデータを共有して使用する環境においても、複数のディスクアレイ 200 間におけるファイ

ルの物理的な再配置を、アプリケーション140、840に対して、再配置前後で論理的に等価となるように行うことができる。

【0149】本実施形態においても、ホスト100のファイルシステム110が、OS120、820やアプリケーション140、840からの各ファイルへのリード/ライト要求頻度を監視し、統計情報を生成して、再配置対象決定処理においてユーザに提示するようにしてもよい。

【0150】本実施形態において、クライアント800上にマネージャ130のプログラムが格納され、そのマネージャ130が、FCインターフェース860あるいはネットワークインターフェース870を用いて、使用状況等の情報の収集や指示などの処理を、ホスト100のファイルシステム110やディスクアレイ200に要求するようにしてもよい。

【0151】図24は、本発明が適用された計算機システムの第4実施形態の構成を示す図である。

【0152】本実施形態の計算機システムは、ホスト100がLUプールマネージャ900及びLU管理テーブル910を有する点で、第1実施形態の計算機システムと異なる。

【0153】本実施形態によれば、LUの再配置先の選択を容易にすることができる。

【0154】図25は、LU管理テーブル910を示す図である。

【0155】LU管理テーブル910は、システム全体のLUの状態に関する情報が登録されているテーブルである。

【0156】LU番号3310には、各LUに一意に割り当てられた番号が登録される。この番号は、LUプールマネージャ900が各LUを管理するために使用される。サイズ3320には、対応するLUの容量が登録される。構成3330には、RAID構成の種別が格納される。構成3330には、LUがキャッシュメモリ330や単体ディスクで構成されている場合には、その情報も格納される。

【0157】状態3340には、LUの状態を示す情報が格納される。その種別として、「オンライン」、「オフライン」、「未実装」及び「障害オフライン」が設けられている。「オンライン」は、LUが正常な状態であり、ホスト100からアクセス可能であることを示す。「オフライン」は空きLU、すなわちLUは正常に存在するが、ホスト100からはアクセス不能の状態におかれていることを示す。「未実装」は、このLUは定義されておらず、ホスト100からアクセス不能であることを示す。「障害オフライン」は、LUに障害が発生してホスト100からのアクセスができないことを示す。

【0158】ディスクアレイ番号3350には、対応するLUが存在するディスクアレイ200を示す情報が格

納される。

【0159】パス3360には、各ディスクアレイ200に複数接続するFC600のどれにLUが割り当てられているかを示す番号が格納される。ID3370及びLUN3380には、LUを示す番号が格納される。

【0160】ディスク性能3390には、対応するLUが現在配置されているディスク装置210の性能を示す指標が格納される。具体的には、図29に示すとおり、ディスク装置210の平均シーク時間、平均回転待ち時間及び構成から、ディスク装置210の性能が高性能、中性能、低性能の指標に分類されて格納されている。キャッシュメモリ上のLUは、超高性能に分類される。

【0161】エミュレーションタイプ3400には、ディスクアレイ200がホスト100に提供する各LUのディスク装置としての型を示す情報が格納される。

【0162】再配置可能フラグ3410には、LUの再配置を行う際に、LUの再配置先として使用できるか否かを指定するためのフラグが格納される。ユーザは、このフラグを用いて再配置用のLUとその他のLUを区別することができる。ユーザはフラグのオン/オフを変更することができる。

【0163】図25は、ディスクアレイ番号0についてのLU管理テーブルを示す図である。マネージャ130は、すべてのディスクアレイ200についてのLU管理テーブルを保持している。

【0164】本実施形態における再配置対象の決定は、以下のように行われる。

【0165】ユーザは、マネージャ130に対して、再配置元LUの指定及び再配置先LUとして必要とされる条件を指定する。具体的な条件としては、性能条件や信頼性レベル等がある。

【0166】例えば、あるLUが過度に使用され、そのLUを含むディスク装置の能力を超えて負荷がかかっている場合、そのLUの再配置先としてより高性能のディスク装置を指定すれば、LUの処理能力が増大し、計算機システムの性能向上が期待できる。

【0167】又、重要なデータを格納しているLUが単体ディスクや冗長なしRAID(RAID0)上に存在する場合、再配置先としてRAID5やRAID1を指定すれば、冗長性による耐障害性を確保できる。

【0168】マネージャ130は、LU管理テーブル910に登録された情報を用いて再配置先のLUを決定し、ユーザに通知した上で、LUの再配置を行う。

【0169】図26は、本実施形態における再配置対象決定処理の手順を示すフロー図である。本処理は、ユーザの指示に対応して実行される。ユーザは、マネージャ130に対して再配置元LUのディスクアレイ番号、パス、ID及びLUNを指定する。この場合、パス及びID等の代わりに、ディスクアレイ番号及びLU番号を指定してもよい(ステップ2500)。

【0170】ユーザは、マネージャ130に対して、再配置先についての要求条件として性能条件や信頼性レベルを指定する(ステップ2510)。

【0171】マネージャ130は、再配置元LU、および再配置先についての要求条件をLUプールマネージャ900に通知する(ステップ2520)。LUプールマネージャ900は、LU管理テーブル910内を検索して、要求された条件を満たすLUの有無を確認する(ステップ2530)。

【0172】この場合、検索条件は、「状態がオフライン」かつ「サイズが再配置元LU以上」かつ「エミュレーションタイプが再配置元LUと同じ」かつ「再配置可能フラグがオン(真)すなわち可能」かつ「性能条件が要求を満たす」かつ「信頼性レベルが要求を満たす」でなければならない。

【0173】ステップ2540において条件を満たすLUが存在した場合、LUプールマネージャは、該当するLUをマネージャ130に通知する(ステップ2550)。マネージャ130は、通知されたLUを再配置先LUとして決定する(ステップ2560)。

【0174】ステップ2540で条件を満たすLUが存在しなかった場合、LUプールマネージャ900は、LU管理テーブル910内を検索して「状態が未実装」のLU番号3310を探す(ステップ2570)。

【0175】未実装のLU番号3310が存在しなかった場合は、LUプールマネージャ900は、マネージャ130に条件を満たすLUの利用不可を通知する(ステップ2580)。通知を受けたマネージャ130は、ユーザに再配置先LU決定不可を通知する(ステップ2590)。

【0176】ステップ2570で未実装のLUが存在した場合は、LUプールマネージャ900は、未実装のLU番号と再配置先LUについての条件を指定して、該当するディスクアレイ200に再配置先LUの構築を指示する(ステップ2600)。

【0177】この場合の再配置先LUについての条件は、「サイズが再配置元LU以上」かつ「エミュレーションタイプが再配置元LUと同じ」かつ「性能条件が要求を満たす」かつ「信頼性レベルが要求を満たす」である。

【0178】LUの構築を指示されたディスクアレイ200は、LU構築処理を行う(ステップ2610)。構築が成功した場合は、ディスクアレイ200は、LUプールマネージャ900に、構築したLUについてのディスクアレイ番号、パス、ID及びLUNなどを含む一連の情報を通知する(ステップ2620)。構築が失敗した場合には、ディスクアレイ200は、LUプールマネージャ900に構築不可の通知を行う(ステップ2610)。

【0179】LUプールマネージャ900は、通知され

たLUの情報をLU管理テーブル910に登録し(ステップ2630)、マネージャ130に通知する(ステップ2550)。マネージャ130は、このLUを再配置先LUとして決定する(ステップ2560)。

【0180】構築不可の通知を受けたLUプールマネージャ900は、マネージャ130に条件を満たすLUの利用不可を通知する(ステップ2580)。通知を受けたマネージャ130は、ユーザに再配置先LU決定不可を通知する(ステップ2590)。

【0181】図27は、ディスクアレイ200が行うLU構築処理の手順を示すフロー図である。この処理は、LUプールマネージャ900の指示を受けた時に行われるディスクアレイ200は、LUプールマネージャ900からの指示により、未実装のLU番号と再配置先LUについての条件を受け取る(ステップ2700)。

【0182】ディスクアレイ200は、ディスク装置210やキャッシュメモリ330などの内部資源割り当て状況等と受け取った条件を比較して、要求された条件のLUが構築可能かどうかを判断する(ステップ2710)。LUが構築可能な場合は、ディスクアレイ200は、内部資源を割り当て、フォーマット/初期化処理を行ってLUを構築する。ディスクアレイ200は、構築したLUに、LUプールマネージャ900から受け取った未実装のLUに対応するLU番号を割り当てる(ステップ2720)。

【0183】ディスクアレイ200は、FCインタフェース260を設定し、LUにパス、ID、LUNを割り当てる(ステップ2730)。ディスクアレイ200は、構築したLUについての、ディスクアレイ番号、パス、ID及びLUN等を含む一連の情報をLUプールマネージャ900に通知する(ステップ2740)。

【0184】ステップ2710においてLUが構築不可能だった場合は、ディスクアレイ200は、構築不可をLUプールマネージャ900に通知する(ステップ2750)。

【0185】再配置先LUが決定されたら、マネージャ130は、第一の実施の形態と同様に再配置元LUと再配置先LUについての再配置処理を行う。

【0186】図28は、再配置元LUのオフライン化処理の手順を示すフロー図である。

【0187】マネージャ130は、第一の実施の形態で説明した方法でコピーの進捗を取得し、コピーが終了した場合は、LUプールマネージャ900に再配置元LUのオフライン化を指示する(ステップ2800)。

【0188】オフライン化の指示を受けたLUプールマネージャ900は、再配置元LUのディスクアレイ200に、再配置元LUのオフライン化を指示する(ステップ2810)。オフライン化の指示を受けたディスクアレイ200は、FCインタフェース260を設定してLUにロックをかけることで、LUをオフラインにする

(ステップ2820)。ディスクアレイ200は、オフライン化したことをLUプールマネージャ900に通知する(ステップ2830)。

【0189】オフライン化の通知を受けたLUプールマネージャは、LU管理テーブル910のLUの状態340の内容をオフラインに更新する(2840)。

【0190】ここでは、マネージャ130がコピーの進捗情報を取得する例を説明したが、ディスクアレイ200がコピー終了をマネージャ130に通知してもよい。

【0191】また、マネージャ130がオフライン化を指示する代わりにディスクアレイ200がコピー終了時点で再配置元LUをオフライン化し、オフライン化したことをLUプールマネージャ900に通知してもよい。

【0192】本実施形態においては、マネージャ130が、SCSI規格のEXTENDED COPYコマンドを用いてスイッチ500のコピー制御部510へのコピー指示を行う場合について説明したが、他のコマンドを用いてもよい。他のコマンドとは、例えば、コマンドポリュームへのライトコマンド等である。また、図15に示すように、ディスクアレイ200がコピー制御部510を有し、マネージャ130がディスクアレイ200のコピー制御部510に、コピー指示を行って、ディスクアレイ200がコピー処理を行うようにしてもよい。

【0193】本実施形態では、再配置先LUとして要求する条件などの情報はユーザが指定しているとしたが、マネージャ130が自動的に判断して指定してもよい。

【0194】本実施形態では、LUプールマネージャ900とマネージャ130が同じホスト100に存在するとしたが、LUプールマネージャ900がリモートコンピュータ400といった、マネージャ130とは異なるコンピュータに存在してもよい。この場合、LUプールマネージャ900とマネージャ130は、FC600やネットワーク700を介してSCSIやSNMPや他のプロトコルやコマンド体系で指示や通知を行う。

【0195】本実施形態によれば、LUの再配置の処理において、再配置先のLUの管理や選択を容易にしてユーザの負荷を削減し、計算機システムの管理を容易にすることができる。

【0196】図29は、本発明を適用した計算機システムの第5実施形態を示す図である。本実施形態の計算機システムは、LU領域範囲テーブル192に新たな項目を付加したLU領域範囲テーブル193を用いて、クライアント800からのリード/ライト要求に基づき、ホスト100がファイルをリード/ライトする。そして、クライアント800との間でネットワーク700を介してデータを転送する処理を行う点が、第3実施形態の計算機システムと異なる。ネットワーク700を経由したファイル共有のためのプロトコルとしては、Network File System (NFS) や Common Internet File System (CI

FS) が広く用いられている。これらのプロトコルや広く普及しているネットワーク700を用いることにより、容易にファイル共有環境を実現することができる。本実施形態においても、NFS又はCIFSを使用することを考える。

【0197】図30は、LU領域範囲テーブル193を示す図である。LU領域範囲テーブル193には、LU領域に対するアクセスがネットワークを使用するか否かに関する情報が格納される。

【0198】使用種別3510には、LU領域が、リード/ライトの処理が第3実施形態のようにFC600を介して行われるLU領域であるか、本実施形態で説明するようにネットワーク700を介してリード/ライトの処理が行われるLU領域であることを示す情報が格納される。

【0199】使用種別3510には、LU領域が、第1実施形態のようにLUを再配置する構成および方法に用いる領域(この場合のリード/ライト要求はFC600を経由する)であるか、LUへのリード/ライトの処理が、ネットワーク700を介して行う領域であるかの情報を格納することができる。使用種別3510には、未使用領域の情報を格納することもできる。その他、領域内アドレス、ディスクアレイ番号、ID、LUN、LU内アドレスは、第3実施形態で説明したものと同様なので、説明を省略する。

【0200】LU領域範囲テーブル193を用いてLUを集中して管理することで、ファイルシステム110は、LUを少なくとも使用種別毎に区別された、複数の領域として管理することができる。

【0201】LU領域範囲テーブル193が設定されることによって、ホスト100は、クライアント800からの要求が、第3実施形態で説明した方法でのアクセスか、ネットワークを介した形でのアクセスかを、要求で用いられるプロトコルなどで区別する。ホスト100は、この種別に応じて、LU領域を使用種別3510毎に区別して扱う。

【0202】ホスト100は、第3実施形態の方法でアクセスされるファイルおよび領域と、本実施形態の方法でアクセスされるファイルおよび領域を区別して処理する。したがって、同一のファイルおよび領域へアクセスする方法が混在することはない。

【0203】ホスト100は、アクセス可能なファイルの検索においても、同様の区別を行う。つまり、同一のディスクアレイ200に存在する各ファイルに、クライアント800からのアクセス要求があった場合、クライアント800からの使用種別を識別することにより、クライアント800の使用種別とは異なる他の使用種別のファイルをクライアント800に対して返答しない。したがって、クライアント800には、自己が使用するアクセス方法でのみアクセス可能なファイルだけが通知さ

れる。このことにより、本システムにおいては、共有ファイルの管理を容易に行うことができる。

【0204】さらに、第1実施形態のように、LUを再配置する構成および方法に用いる領域（リード／ライトはFC600経由）と、LUへのリード／ライトをホスト100およびネットワーク700を介して行う領域との区別を行うことによって、上述したような効果をこれら全ての使用種別に対して得ることができる。又、ユーザは、ホスト100またはリモートコンピュータ400を介して、LU領域範囲テーブル193を自由に設定することができる。

【0205】本実施形態では、NFSやCIFSのようなファイル共有プロトコルをネットワーク700経由で使用し、ホスト100とクライアント800間のデータ転送をネットワーク700経由で行うとしたが、ネットワーク700の代わりに、FC800経由で行う処理も考えられる。更に、クライアント800が行う各LUへのリード／ライト要求を、ホスト100およびネットワーク700を介して行う処理も考えられる。この場合、ホスト100は、クライアント800が要求するリード／ライト対象領域を、LU領域範囲テーブル192を用いて求める。ホスト100は、対象となるデータをリードしてクライアント800にネットワーク700経由で転送する。あるいは、ホスト100は、クライアント800からデータをネットワーク700経由で受領してライトする。

【0206】図31は、クライアント800のアプリケーション840が、ディスクアレイ200に格納されているファイルに対してリードを行う場合における、ホスト100の処理の手順を示すフロー図である。

【0207】第3実施形態と同様に、リード通知を受けたホスト100のファイルシステム110は、LU領域範囲テーブル193とメタデータを参照することで、ファイルの格納されているLU及びLU内領域を求める（ステップ2900）。ファイルシステム110は、他のライト要求に対してリード対象となるファイルをロックする（ステップ2910）。ファイルシステム110は、ファイル内のデータをリードして（ステップ2920）、クライアント800にネットワーク700を介してリードした内容を転送する（ステップ2930）。ファイルシステム110は、メタデータ上のファイルアクセス日時を更新する（ステップ2940）。ファイルシステム110は、ファイルのロックを解除し（ステップ2950）、リード処理の完了をクライアント800に通知する（ステップ2960）。

【0208】図32は、アプリケーション840が、ライトを行う場合の処理の手順を示すフロー図である。

【0209】ライト通知を受けたホスト100は、ネットワーク700経由でクライアント800からライトデータを受け取る（ステップ3000）。ホスト100

は、LU領域範囲テーブル193とメタデータを参照することで、ファイルの格納されているLUとLU内領域を求める（ステップ3010）。ホスト100は、ファイルをロックし（ステップ3020）、ファイルにデータをライトする。このとき、必要ならばメタデータを更新してファイル使用領域の追加を行う（ステップ3030）。

【0210】ホスト100は、メタデータ上のファイル更新日時とアクセス日時を更新する（ステップ3040）。ホスト100は、ロックを解除し（ステップ3050）、ライト完了をクライアント800に通知する（ステップ3060）。

【0211】図33は、クライアント800のアプリケーション840またはOS820が、アクセス可能なファイルの存在についてホスト100に問い合わせた場合の処理の手順を示すフロー図である。

【0212】アプリケーション840またはOS820自身の要求があった時、OS820は、ネットワーク700を介してホスト100にアクセス可能なファイルの存在を問い合わせる（ステップ3100）。

【0213】通知を受けたホスト100のファイルシステム110は、アクセス可能なファイルを、LU領域範囲テーブル193とメタデータを参照して求める（ステップ3110）。ファイルシステム110は、各ファイルのファイル名などの情報をクライアント800に通知する（ステップ3120）。

【0214】本実施形態では、クライアント800およびホスト100は、ディスクアレイ200に格納されているファイルをホスト100経由で共有して使用することができる。データの再配置の方法等は第3実施形態と同様である。ただし、再配置処理は各使用種別の領域内で行われる。

【0215】本実施形態では、ディスクアレイ200に格納されているデータを共有して使用する環境においても、アプリケーション140およびアプリケーション840が関与することなく、複数のディスクアレイ200間でのファイルの物理的再配置を行うことができる。

【0216】本発明は、各実施形態に限定されるものではなく、その要旨の範囲内で数々の変形が可能である。

【0217】たとえば、図1、図15および図21に示すように、マネージャ130を、ネットワークインタフェース470とFCインタフェース460とを有するリモートコンピュータ400上のプログラムとして、ホスト100の外部に配置してもよい。ホスト100外部のマネージャ130がFC600あるいはネットワーク700経由で情報の収集や指示を行い、各実施形態と同様の処理を行って、複数のディスクアレイ200間におけるLUの再配置によるデータの適正配置を、アプリケーション140に対して透過的に行うことができる。

【0218】また、第1実施形態において、第3実施形

態と同様に、ファイルの共有等を行うようにしてもよい。この場合も、記第1実施形態と同様に、複数のディスクアレイ200間でのデータの物理的再配置を、アプリケーション140、840に対し、再配置前後で論理的に等価となるように透過的に行うことができる。

【0219】

【発明の効果】以上説明したように、本発明によれば、ストレージサブシステム間におけるデータの再配置を容易に行うことが出来る。また、本発明によれば、ホストコンピュータが適正配置の決定に必要な情報を複数のストレージサブシステムから取得することができる。また、異なるストレージサブシステム間におけるデータの再配置を、アプリケーションにとってのデータ位置が、再配置の前後で変化しないように行うことができる。さらに、異なるストレージサブシステム間におけるファイルを対象とするデータの再配置を行うことができる。

【図面の簡単な説明】

【図1】本発明の第1実施形態が適用された計算機システムの構成を示す図である。

【図2】本発明の第1実施形態でのリード/ライト処理および使用状況取得処理の手順を示すフロー図である。

【図3】本発明の第1実施形態で用いる論理/物理対応情報を示す図である。

【図4】本発明の第1実施形態で用いる論理ボリューム使用状況を示す図である。

【図5】本発明の第1実施形態での使用状況収集処理の手順を示すフロー図である。

【図6】本発明の第1実施形態で用いる論理ボリューム情報のパラメータを示す図である。

【図7】本発明の第1実施形態で用いる論理ボリューム情報を示す図である。

【図8】本発明の第1実施形態での再配置対象決定処理の手順を示すフロー図である。

【図9】本発明の第1実施形態での再配置処理の手順を示すフロー図である。

【図10】本発明の第1実施形態で用いるLU論理位置名テーブルを示す図である。

【図11】本発明の第1実施形態で用いるLU論理位置名テーブルを示す図である。

【図12】本発明の第1実施形態でのコピー処理の手順を示すフロー図である。

【図13】本発明の第1実施形態で用いるコピー領域管理テーブルを示す図である。

【図14】本発明の第1実施形態での、図16に示す処理によるコピー中における、再配置先LUへのリード/ライトコマンドに対する処理の手順を示すフロー図である。

【図15】本発明の第2実施形態が適用された計算機システムの構成を示す図である。

【図16】本発明の第2実施形態で用いるLU領域範囲テ

ーブルを示す図である。

【図17】本発明の第2実施形態でのリード/ライト処理の手順を示すフロー図である。

【図18】本発明の第2実施形態で用いるメタデータを示す図である。

【図19】本発明の第2実施形態での再配置対象決定処理の手順を示すフロー図である。

【図20】本発明の第2実施形態での再配置処理の手順を示すフロー図である。

【図21】本発明の第3実施形態が適用された計算機システムの構成を示す図である。

【図22】本発明の第3実施形態において、クライアントのアプリケーションがファイルのリードを行う際の処理の手順を示すフロー図である。

【図23】本発明の第3実施形態において、クライアントのアプリケーションがファイルのライトを行う際の処理の手順を示すフロー図である。

【図24】本発明の第4実施形態が適用された計算機システムの構成を示す図である。

【図25】本発明の第4実施形態でのLU管理テーブル910を示す図である。

【図26】本発明の第4実施形態での再配置対象決定処理の手順を示すフロー図である。

【図27】本発明の第4実施形態でのLU構築処理の手順を示すフロー図である。

【図28】本発明の第4実施形態での再配置元LUオフライン化処理の手順を示すフロー図である。

【図29】本発明の第5実施形態が適用された計算機システムの構成を示す図である。

【図30】本発明の第5実施形態でのLU領域範囲テーブル193を示す図である。

【図31】本発明の第5実施形態でのクライアント800のアプリケーション840がファイルのリードを行う際の処理の手順を示したフロー図である。

【図32】本発明の第5実施形態でのクライアント800のアプリケーション840がファイルのライトを行う際の処理の手順を示すフロー図である。

【図33】本発明の第5実施形態でのアクセス可能ファイル応答処理の手順を示すフロー図である。

【符号の説明】

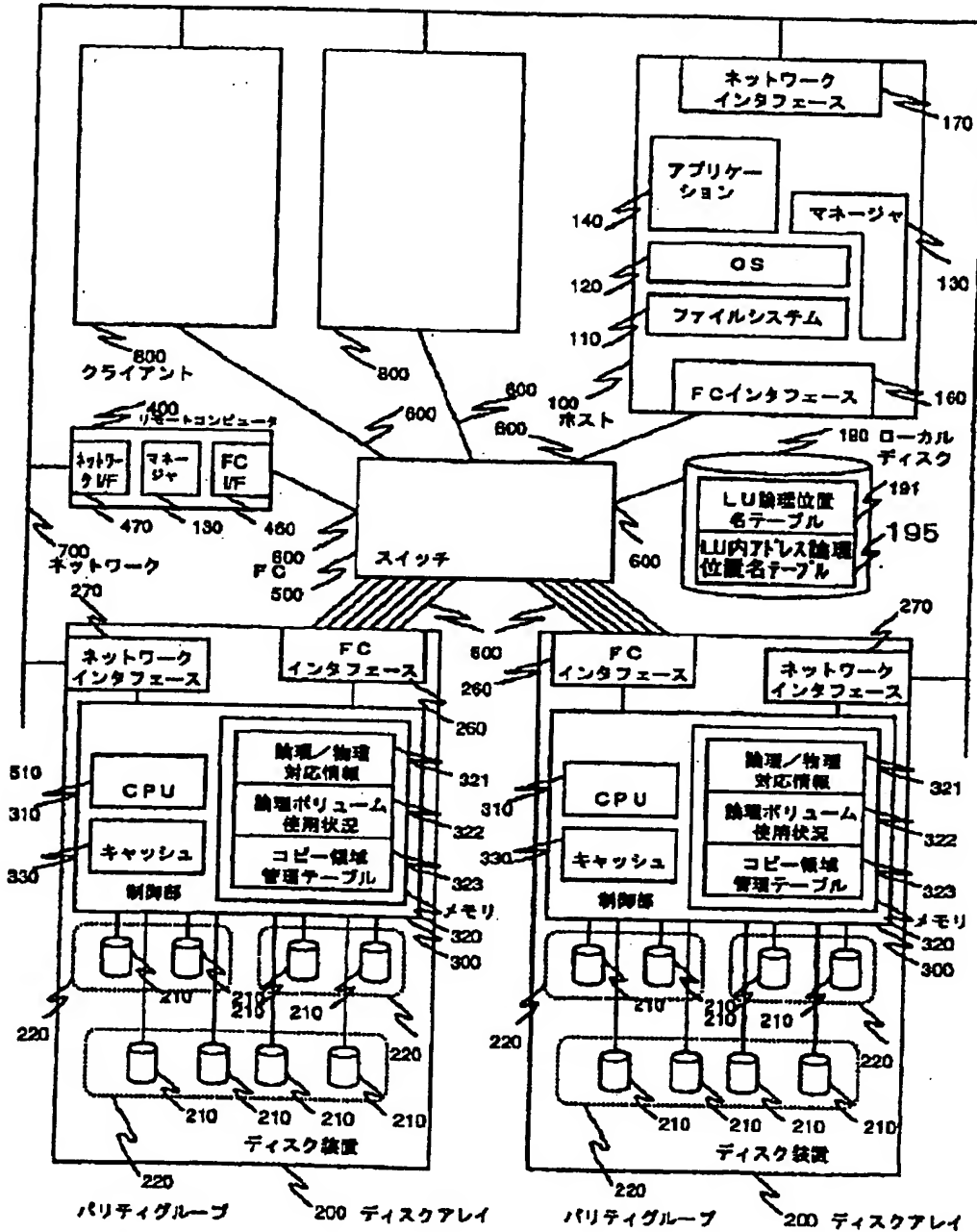
100…ホスト、110…ファイルシステム、120、820…OS、130…マネージャ、140、840…アプリケーション、160、260、460、860…FCインタフェース、170、270、470、870…ネットワークインタフェース、190…ローカルディスク、191…LU論理位置名テーブル、192…LU領域範囲テーブル、200…ディスクアレイ、210…ディスク装置、220…パリティグループ、300…制御部、310…CPU、320…メモリ、321…論理/物理対応情報、322…論理ボリューム使用状況、323…コピ

一領域管理テーブル、330…キャッシュメモリ、400…リモートコンピュータ、500…スイッチ、510

…コピー制御部、600…Fibre Channel (FC)、700…ネットワーク、800…クライアント。

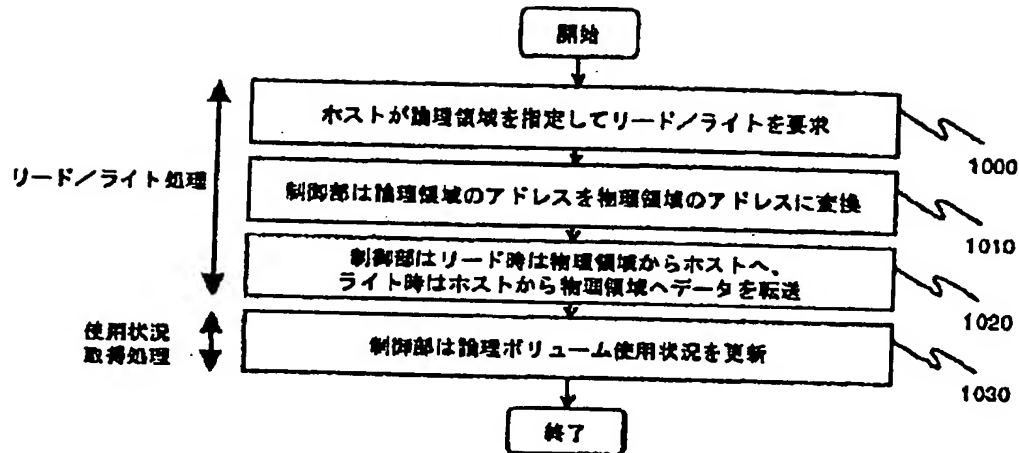
【図1】

図1



【図2】

図2



【図3】

図3

論理アドレス				物理アドレス				
5001	5002	5003	5004	5005	5006	5007	5008	5009
LU 番号	LU内 アドレス	論理ボ リューム 番号	論理ボ リューム アドレス	パリティ グループ 番号	ディスク 装置 番号	ディスク 装置内 アドレス	ディスク 装置 番号	ディスク 装置内 アドレス
0	0~999	0	0~999	100	0	0~999	2	0~999
0	1000~1999	0	1000~1999	100	1	0~999	2	0~999
0	2000~2999	1	0~999	100	0	1000~1999	2	1000~1999
0	3000~3999	1	1000~1999	100	1	1000~1999	2	1000~1999
0	4000~4999	2	0~999	101	3	0~999	4	0~999
1	0~999	3	0~999	100	5	0~999	5	0~999
2	0~999	4	0~999	コマンドボ リューム	-	-	-	-

【図11】

図11

LU番号			論理位置名(ドライブ)
ディスクアレイ 番号	ID	LUN	
0	2	5	X:
1	4	5	Y:
1	5	2	使用していない(空)

【図13】

図13

6101	6102	6103	6104
コピー元 LU番号	コピー元 LU番号	コピー ブロック数	ビットマップ
11:22:33:44: 55:66:77:88	10:20:30:40: 50:60:70:80	1000	1 1 1 1 1 1 1 1 1 1

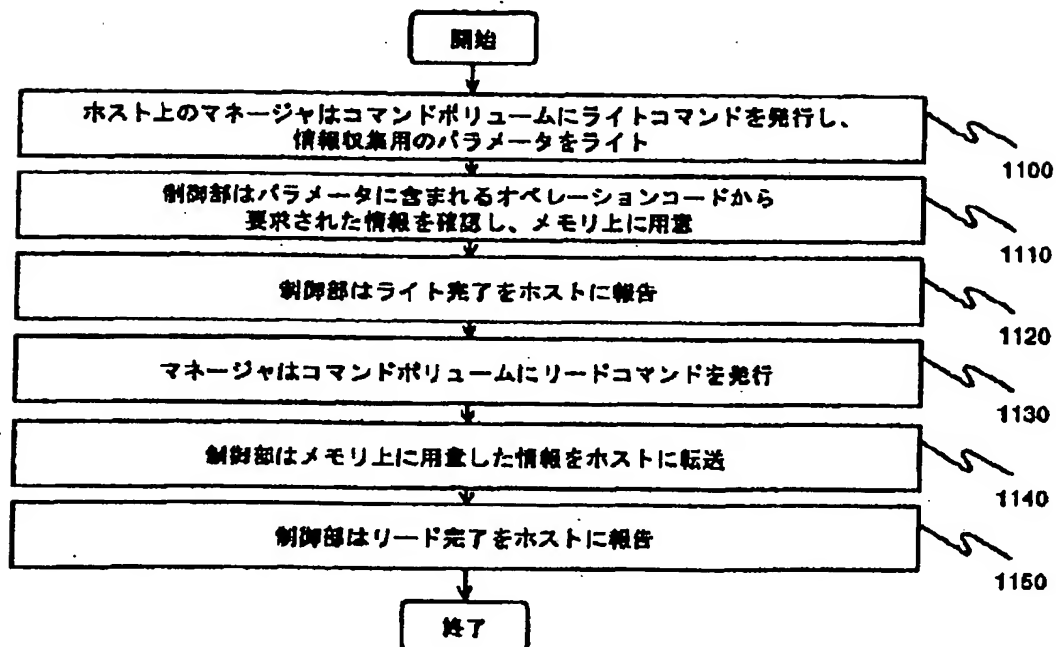
【図4】

図4

論理ボリューム番号	5102					
	ランダムリード処理によるパリティグループ占有時間(マイクロ秒単位)	ランダムライト(データ)処理によるパリティグループ占有時間(マイクロ秒単位)	ランダムライト(パリティ)処理によるパリティグループ占有時間(マイクロ秒単位)	シーケンシャルリード処理によるパリティグループ占有時間(マイクロ秒単位)	シーケンシャルライト(データ)処理によるパリティグループ占有時間(マイクロ秒単位)	シーケンシャルライト(パリティ)処理によるパリティグループ占有時間(マイクロ秒単位)
0	3141	101	121	59265	0	0
1	0	131	111	0	24380	17320

【図5】

図5



【図6】

図6

バイト	内容	
0-1	論理ボリューム番号	5201
2-4	Reserved	
5	0xB6:オペレーションコード(論理ボリューム情報取得)	
6-7	Reserved	
8-511	Don't care	

【図7】

図7

バイト	内容	
0-1	論理ボリューム番号	5201
2-4	Don't care	
5	0xB6:オペレーションコード(論理ボリューム情報取得)	5202
6-7	Don't care	
8-11	ディスクアレイ製造番号	
12-43	ディスクアレイがホストに対しエミュレーションしてみせるボリューム形式 (ボリュームエミュレーションタイプ)(コードです)	5203
44-47	ボリューム容量(MB単位)	
48	ボリュームがコピー先となるか否かを表すフラグ	
49	当該LUを構成する論理ボリューム数	
51-121	当該LUを構成する論理ボリューム番号(2バイト毎)	5203
122-511	Reserved	

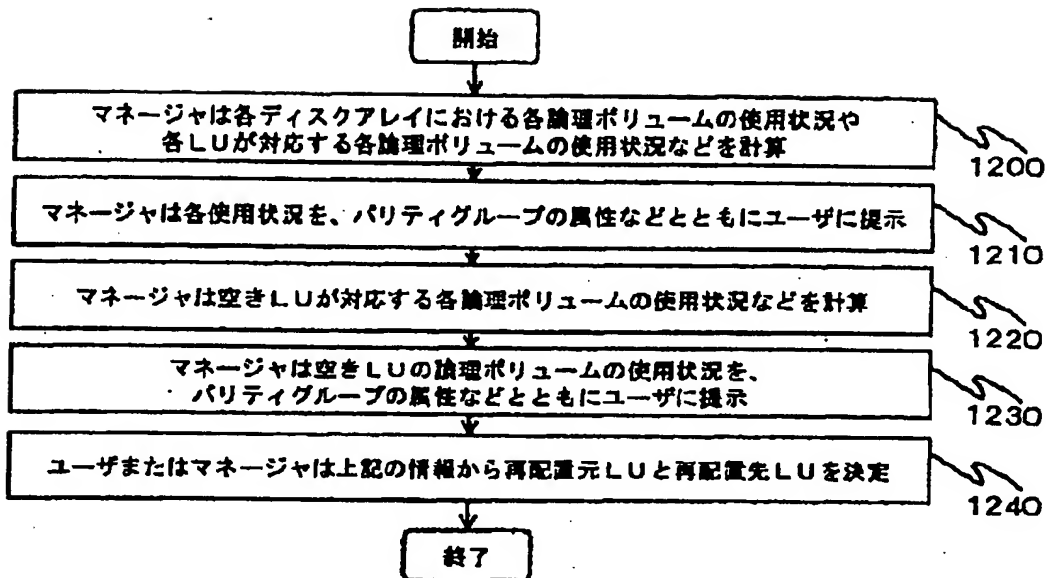
【図10】

図10

LU番号			論理位置名(ディレクトリ)
ディスクアレイ 番号	ID	LUN	
0	2	3	.../home1/prjct1
1	4	5	.../temp
1	5	2	使用していない(空き)

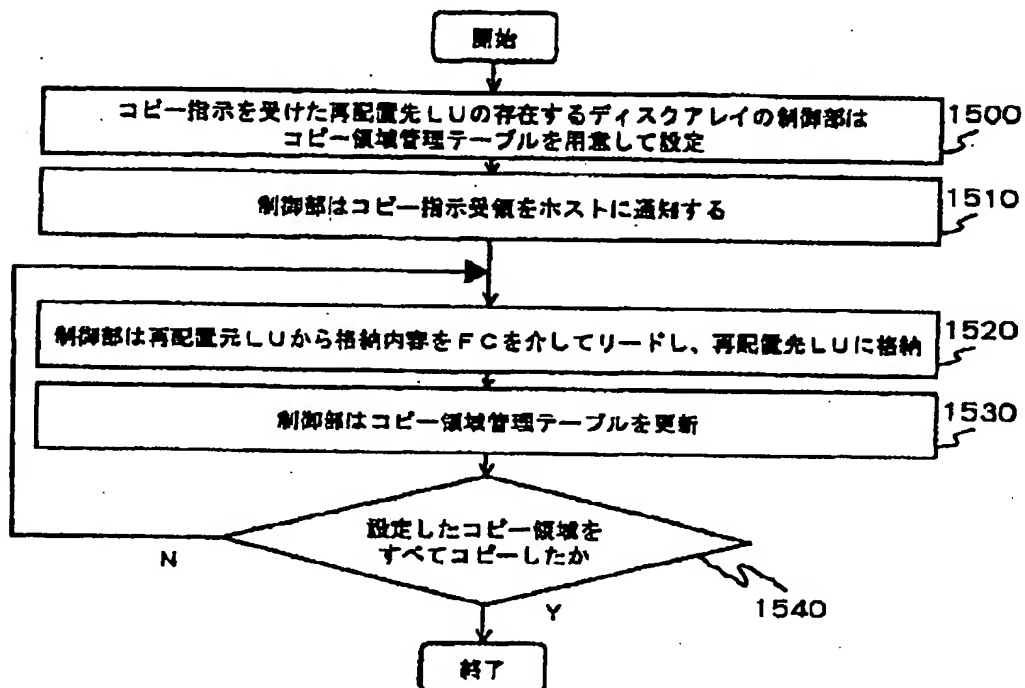
【図8】

図8



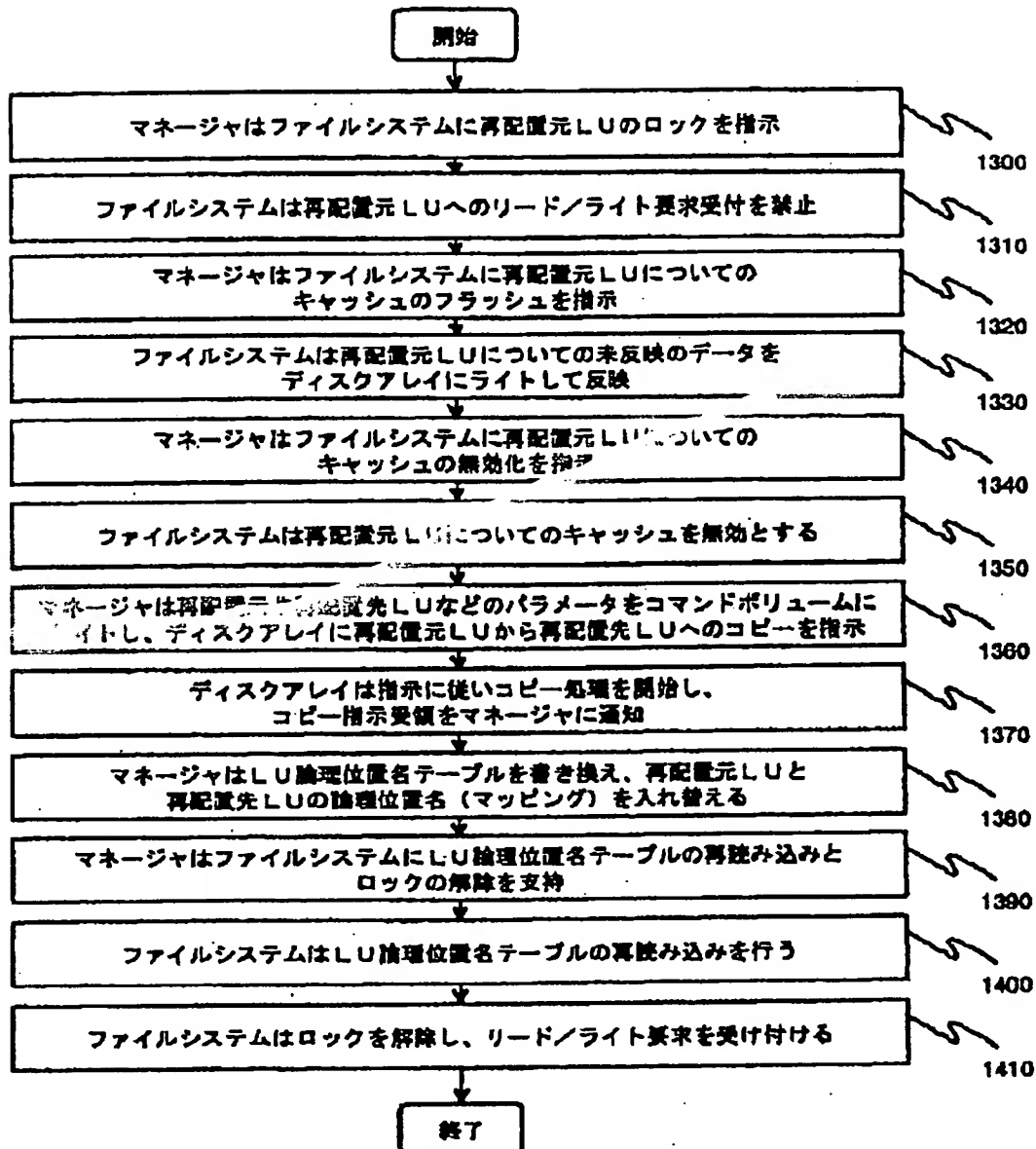
【図12】

図12

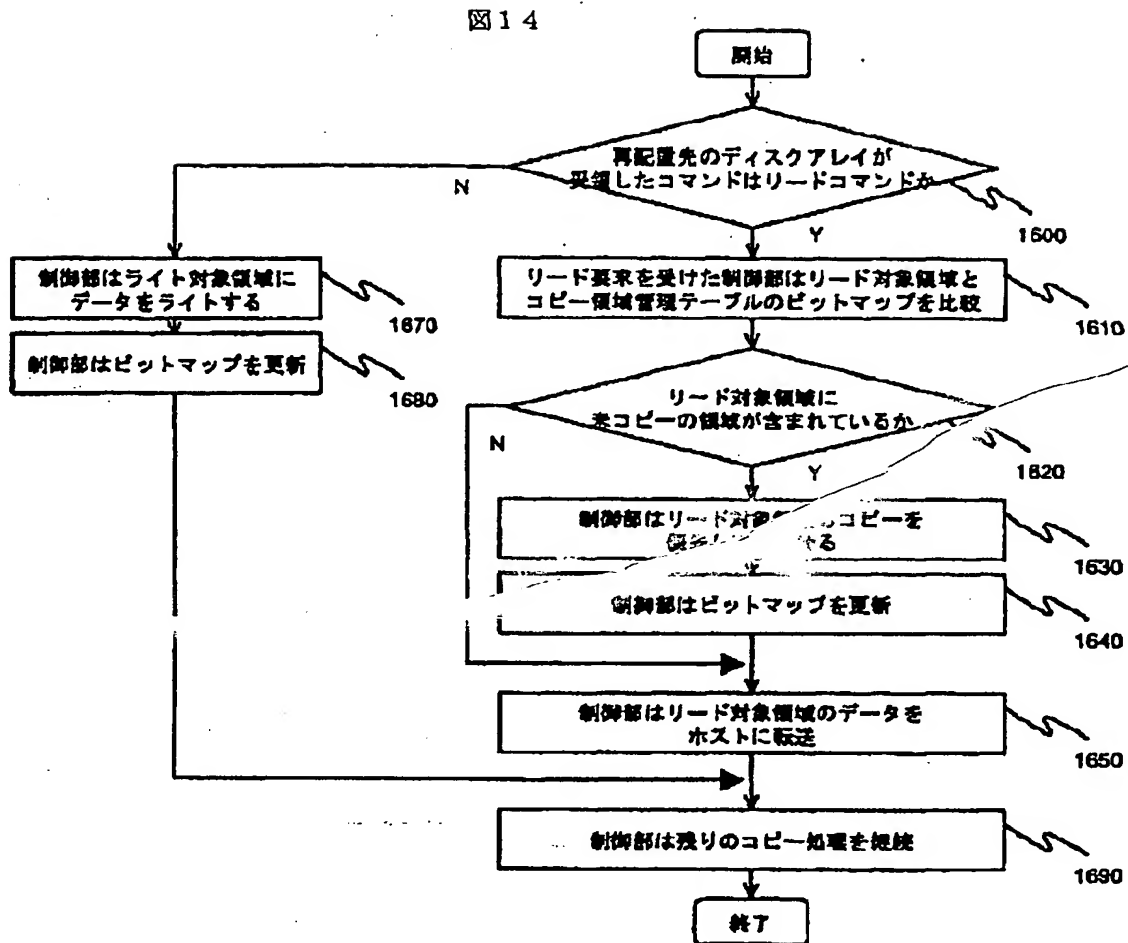


【図9】

図9



【図14】



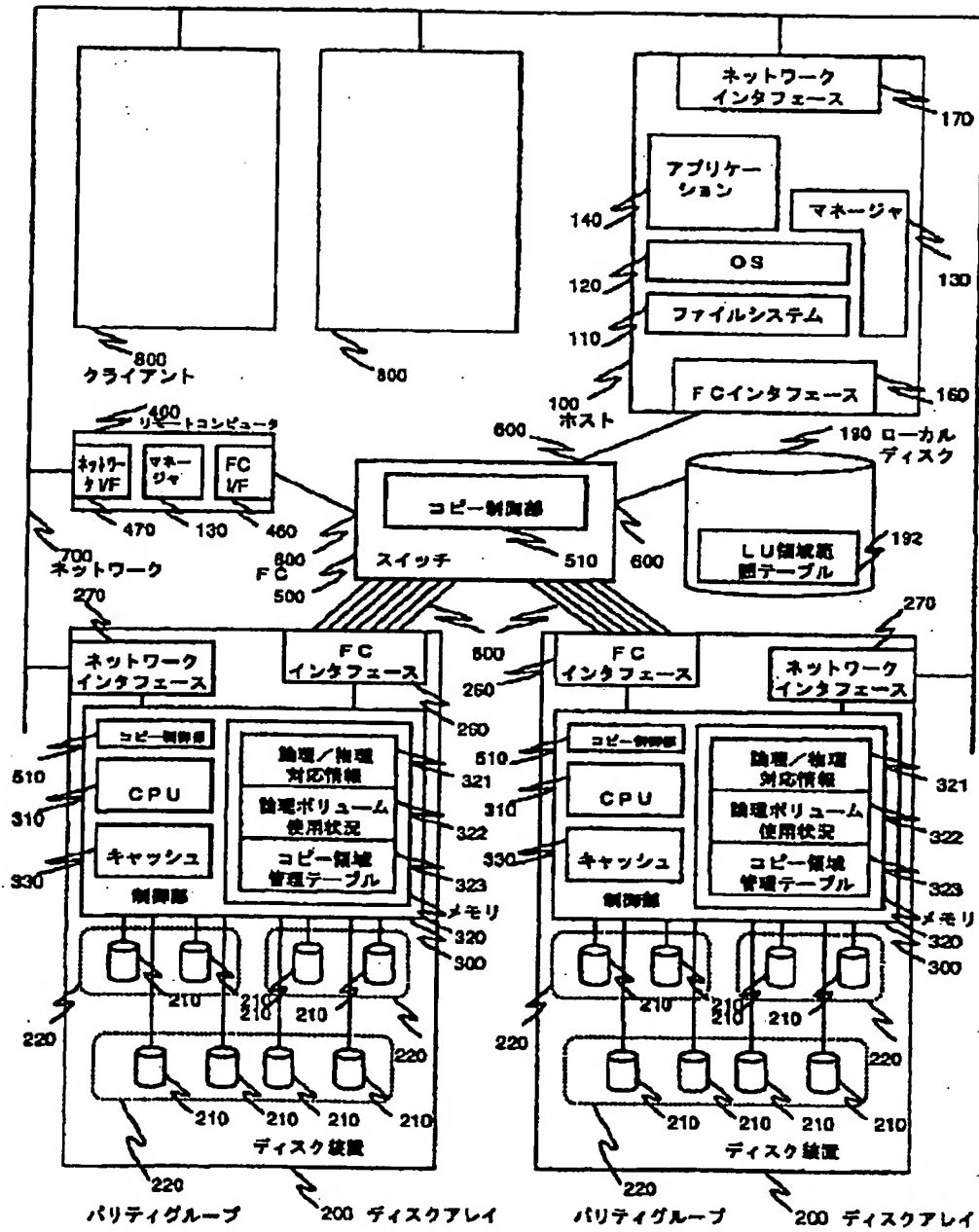
【図16】

図16

領域内アドレス	LU番号			LU内アドレス
	ディスクアレイ 番号	ID	LUN	
0~999	0	2	3	0~999
1000~2999	1	4	5	0~1999
3000~3999	1	5	2	0~999

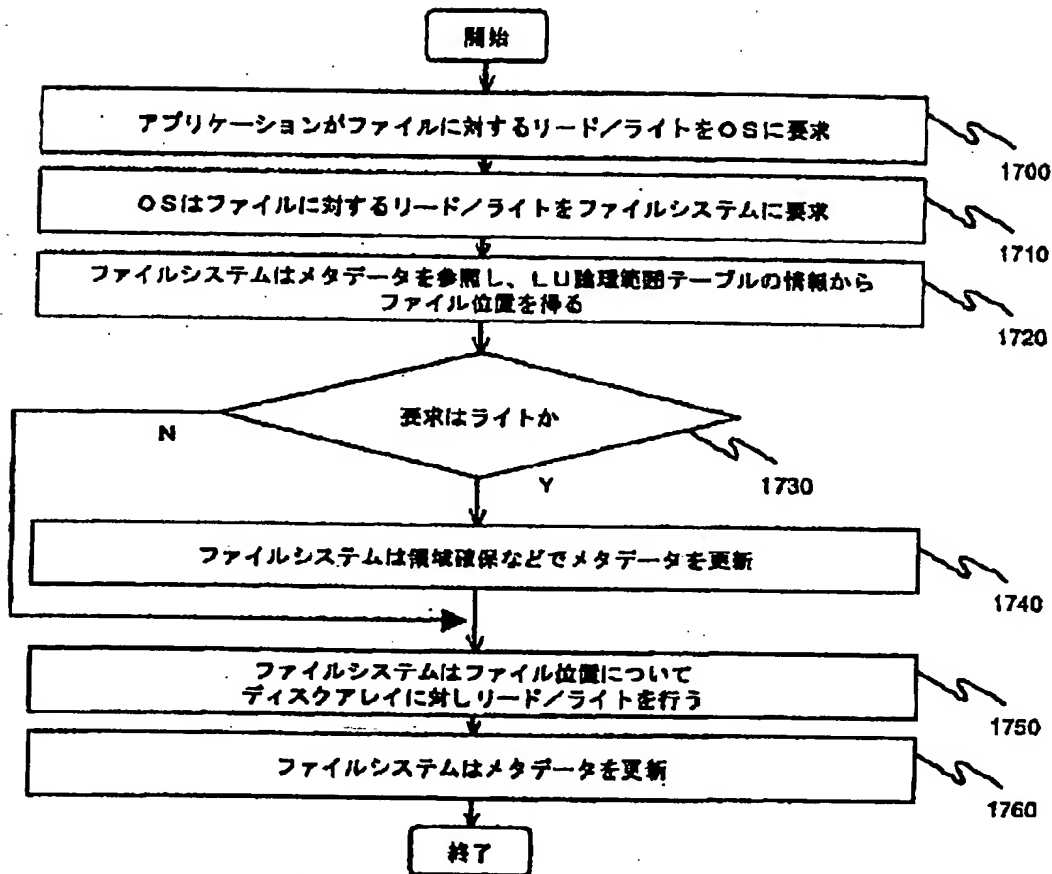
【図 15】

図 15



【図17】

図17



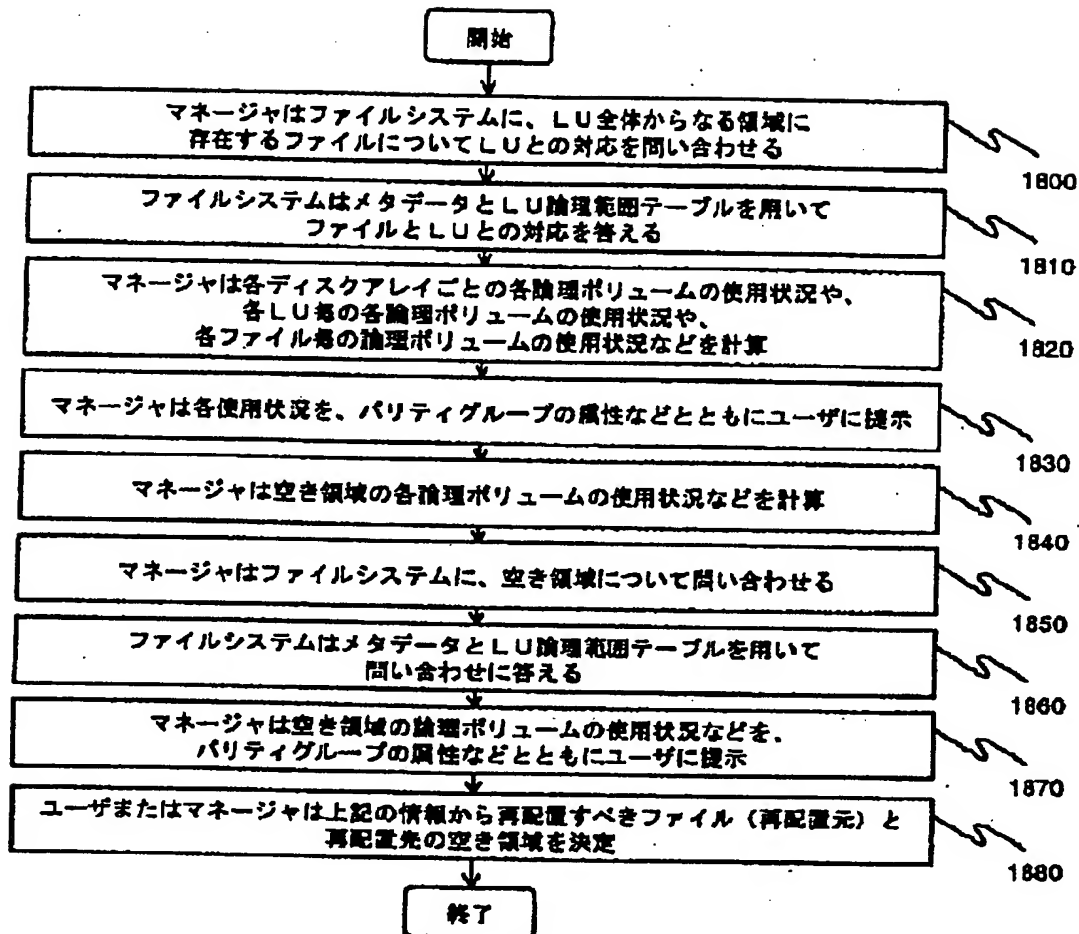
【図18】

図18

6401 作成日時	6402 更新日時	6403 アクセス日時	6404 属性	6405 ファイル 論理位置名	6406 セキュリティ 情報	6407 ファイル 位置
Jan 7, 2000 19:30	Jan 7, 2000 19:40	Jan 7, 2000 19:50	ノーマル	.../prJct1 /doc1.txt	オーナーのみ のみリード /ライト可	100
Jan 9, 2000 8:30	Jan 9, 2000 8:40	Jan 10, 2000 13:50	システム ファイル	.../prJct3 /voice1.mld	グループの のみリード/ ライト可	300

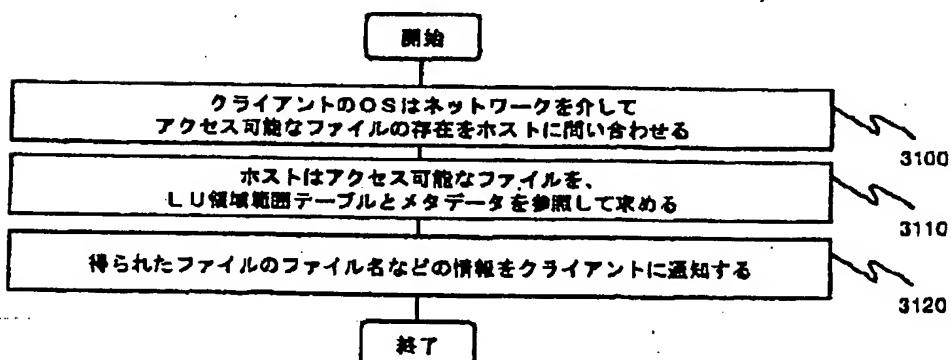
【図19】

図19



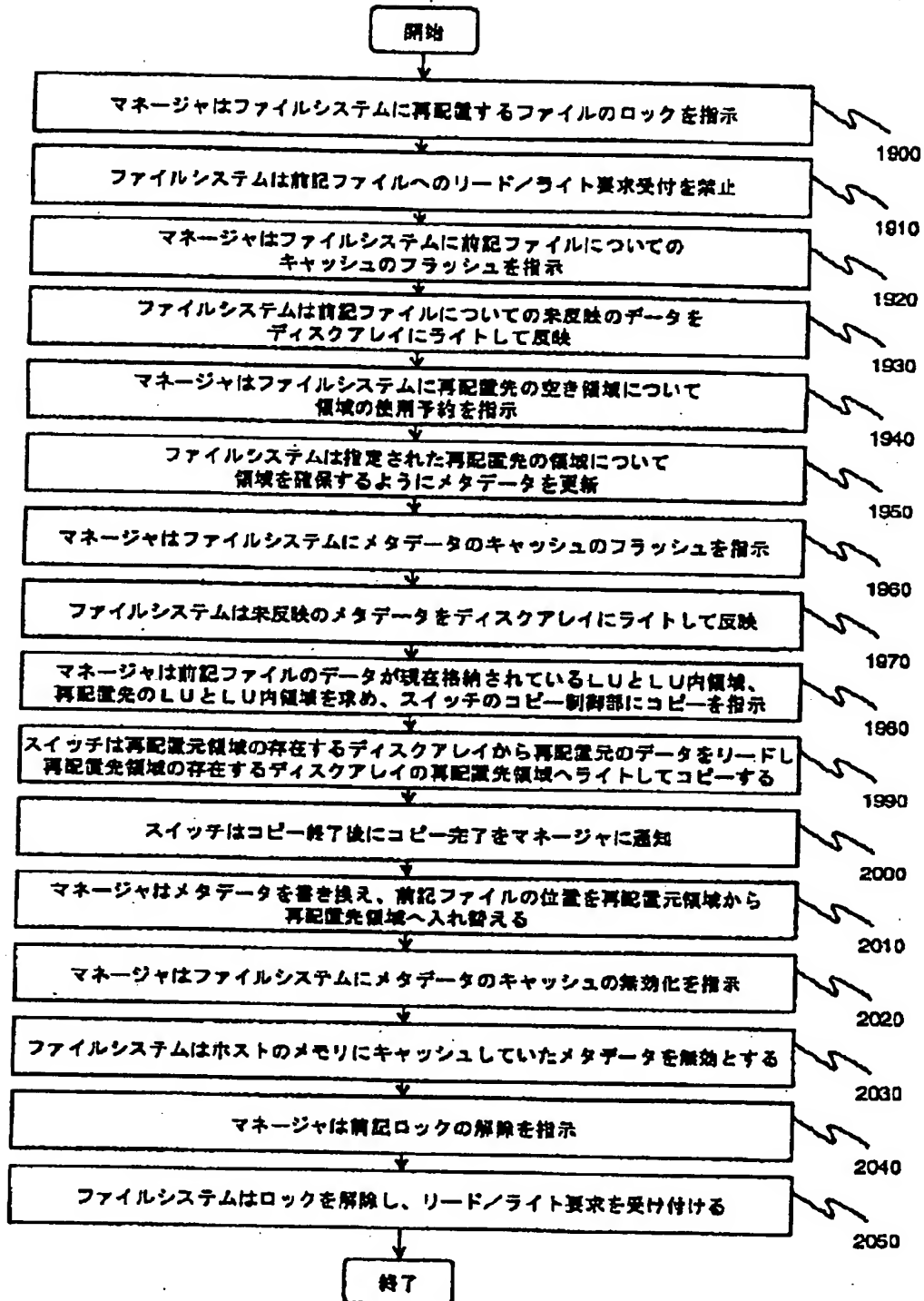
【図33】

図33



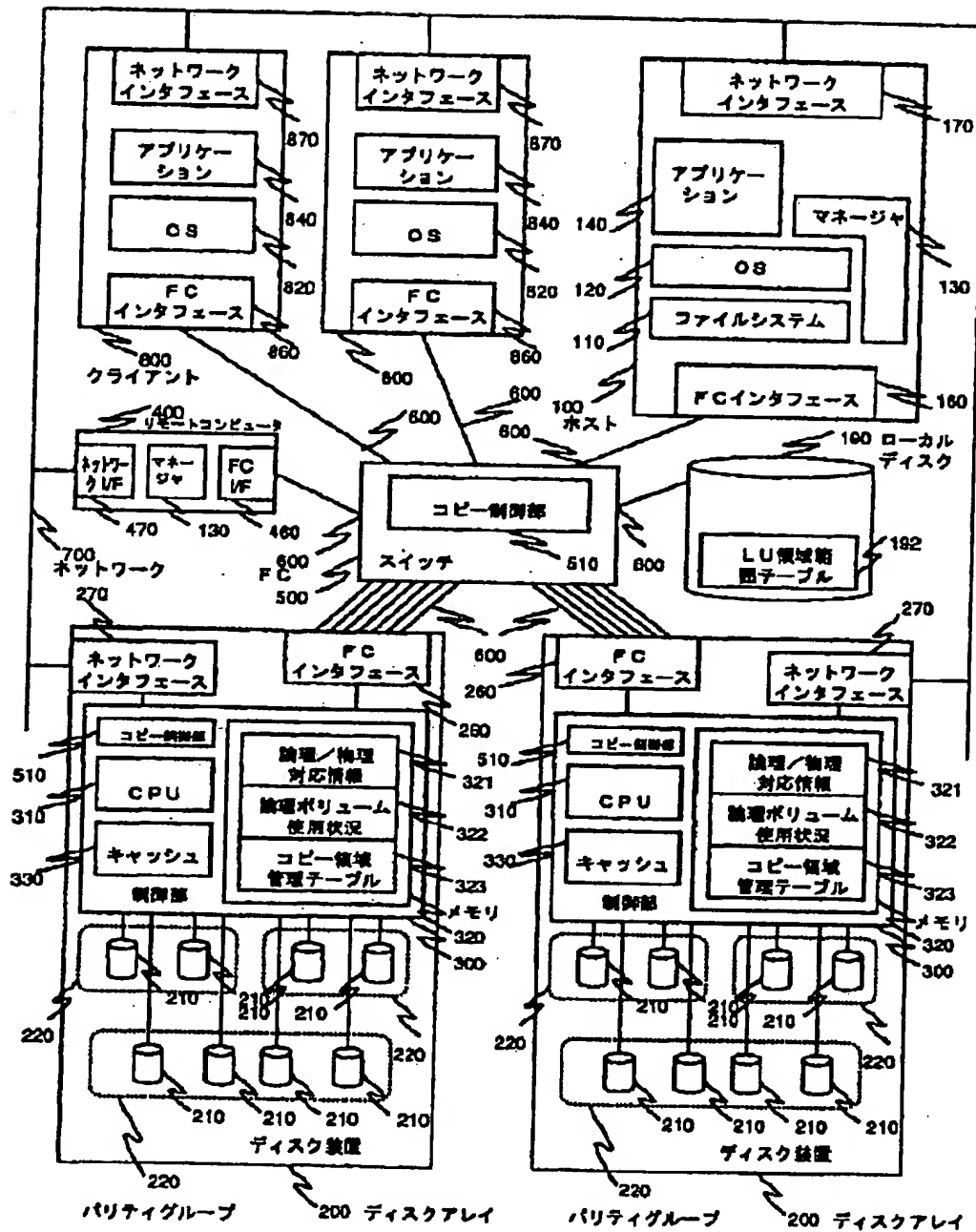
【図20】

図20



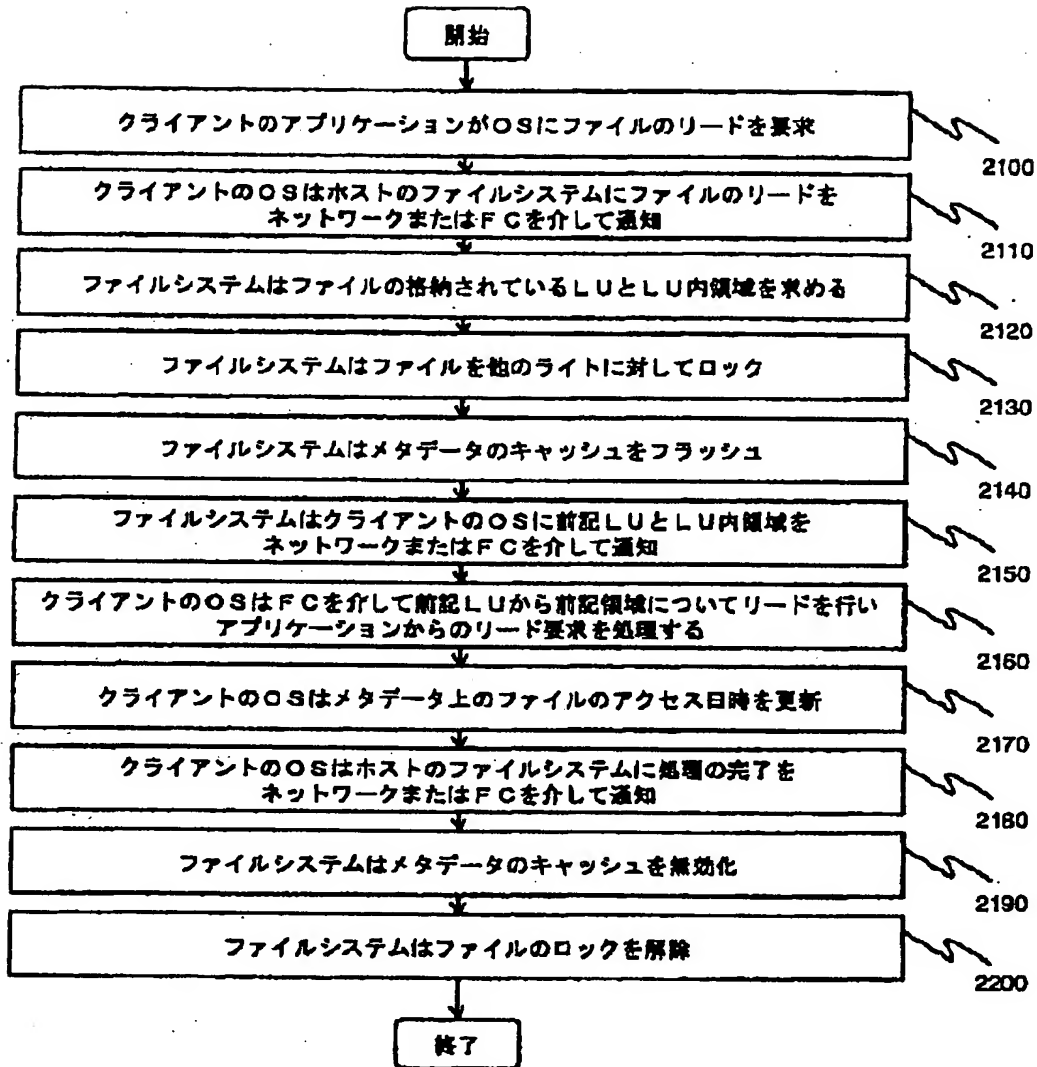
【図21】

図21



【図22】

図22



【図23】

図23

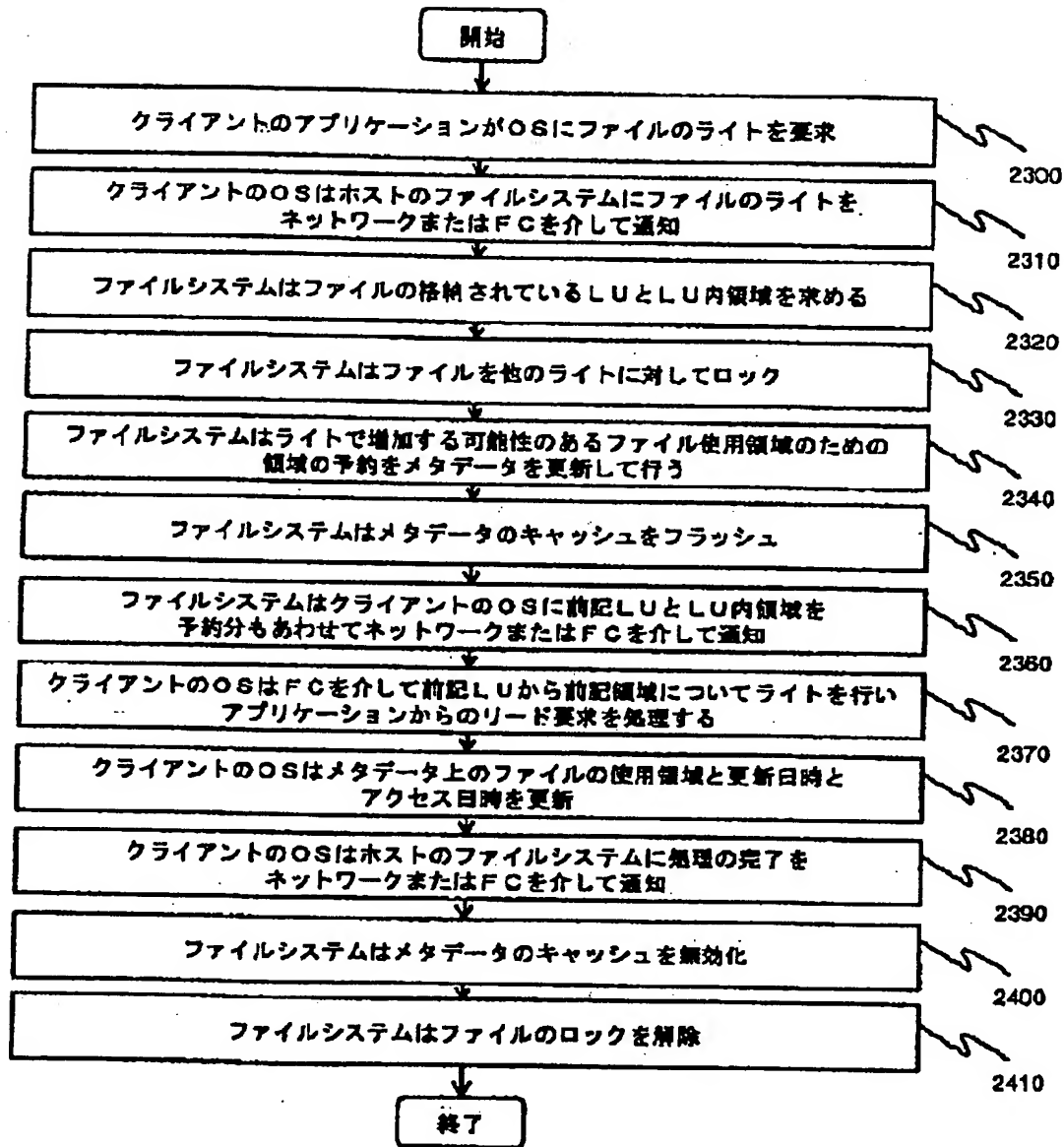
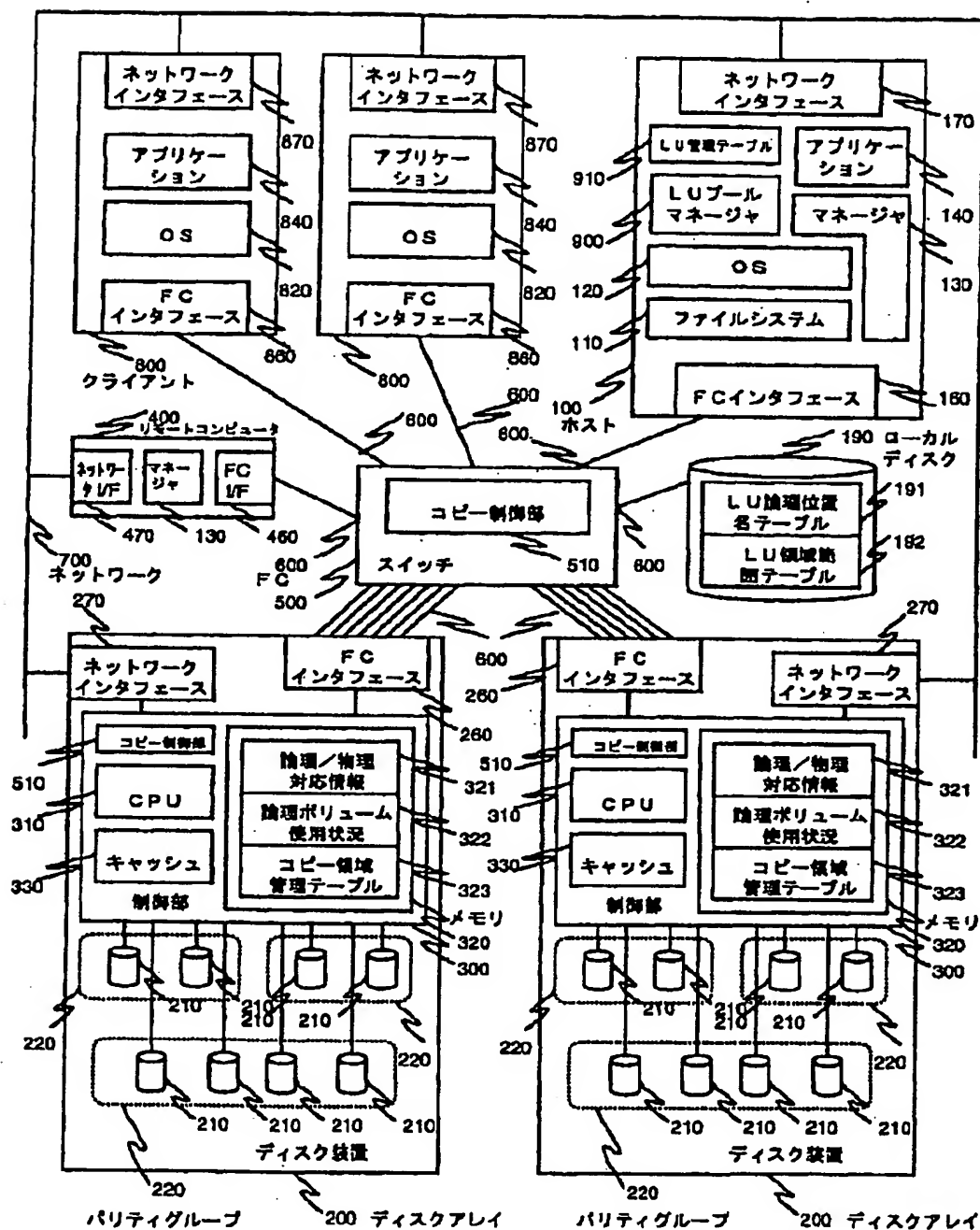


图 24



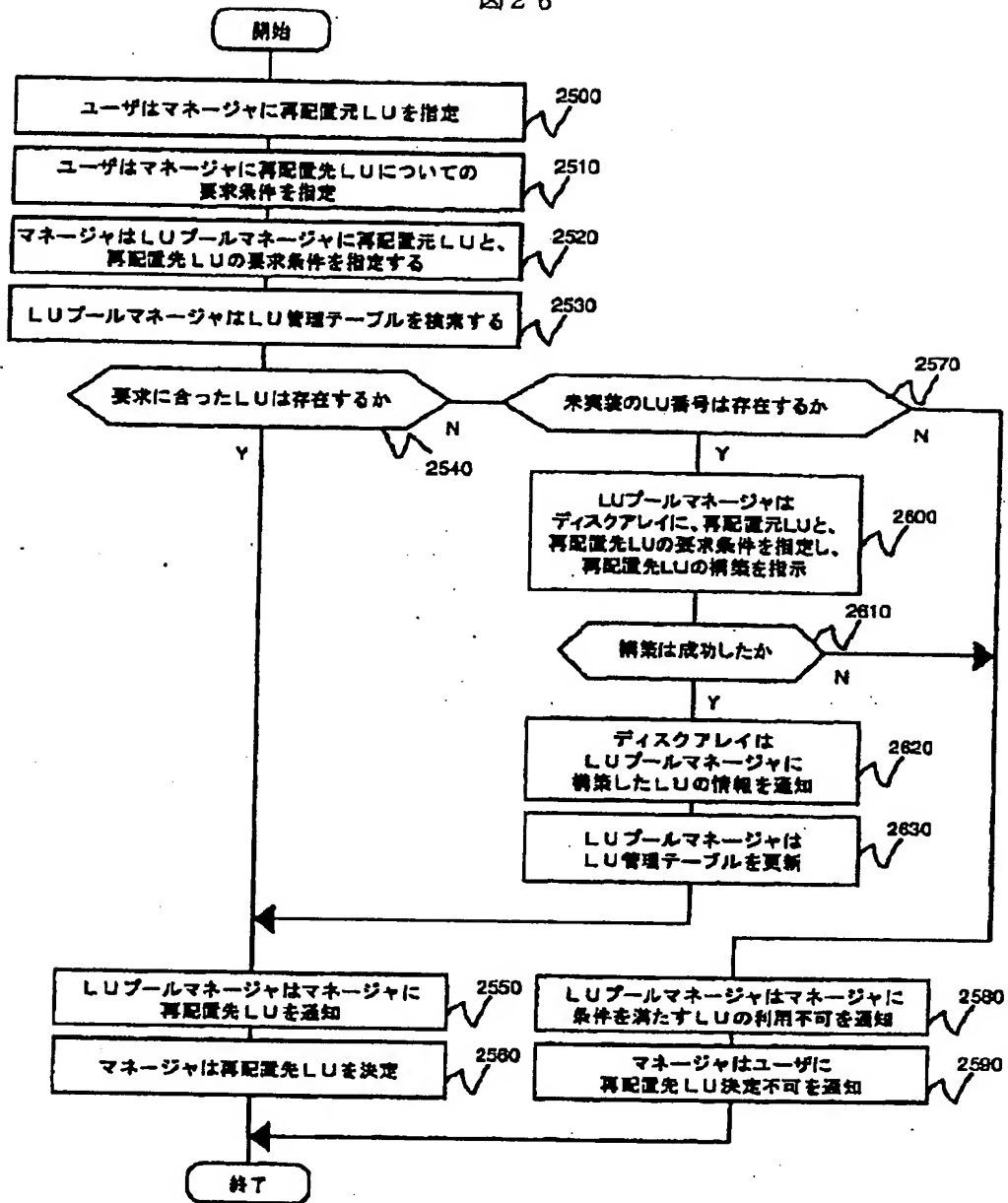
【図25】

図25

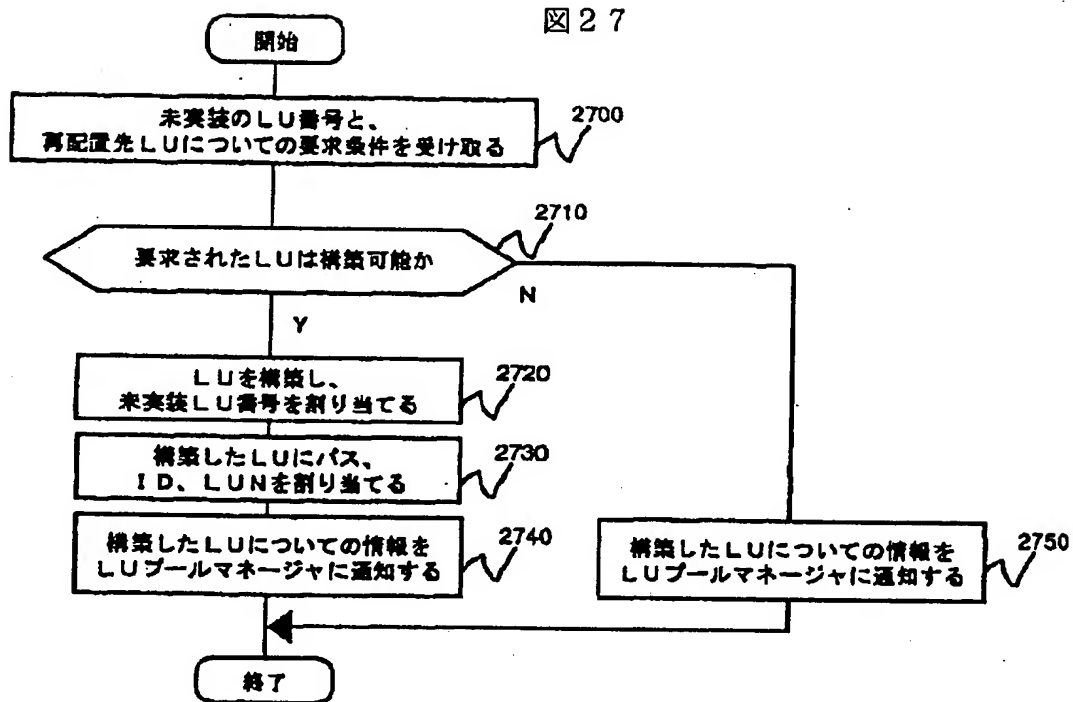
LU 番号	サイズ	構成	状態	ディスク アレイ 番号	バス	ID	L U N	ディスク 性能	エミュレー ション タイプ	再配置 可能 フラグ
0	1000	RAID1	オンラ イン	0	0	0	0	中性能	Open3	オフ
1	1000	RAID5	オフラ イン	0	-	-	-	高性能	Open3	オン
2	3500	キャッシュ	オンラ イン	0	0	0	1	超高性能	Open3	オフ
2	3500	キャッシュ	オンラ イン	0	1	0	0	超高性能	Open3	オフ
k	-	-	未実 装	-	-	-	-	-	-	-
k+1	1000	単体デ ィスク	障害 オフラ イン	0	0	1	0	低性能	3390-3	オフ
k+2	1000	RAID1	オフラ イン	0	-	-	-	高性能	3390-3	オン
n	-	-	未実 装	-	-	-	-	-	-	-

【図26】

図26

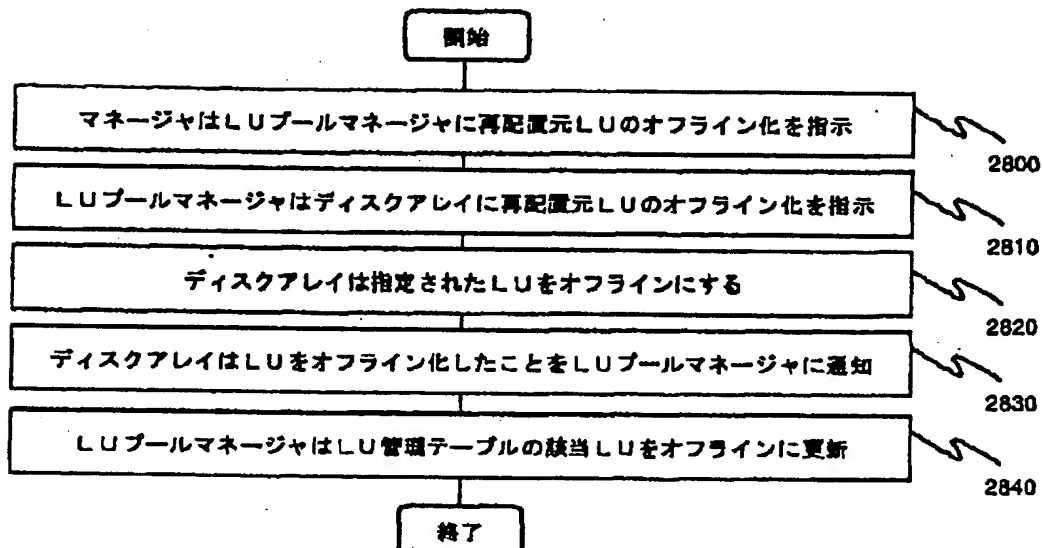


【図27】



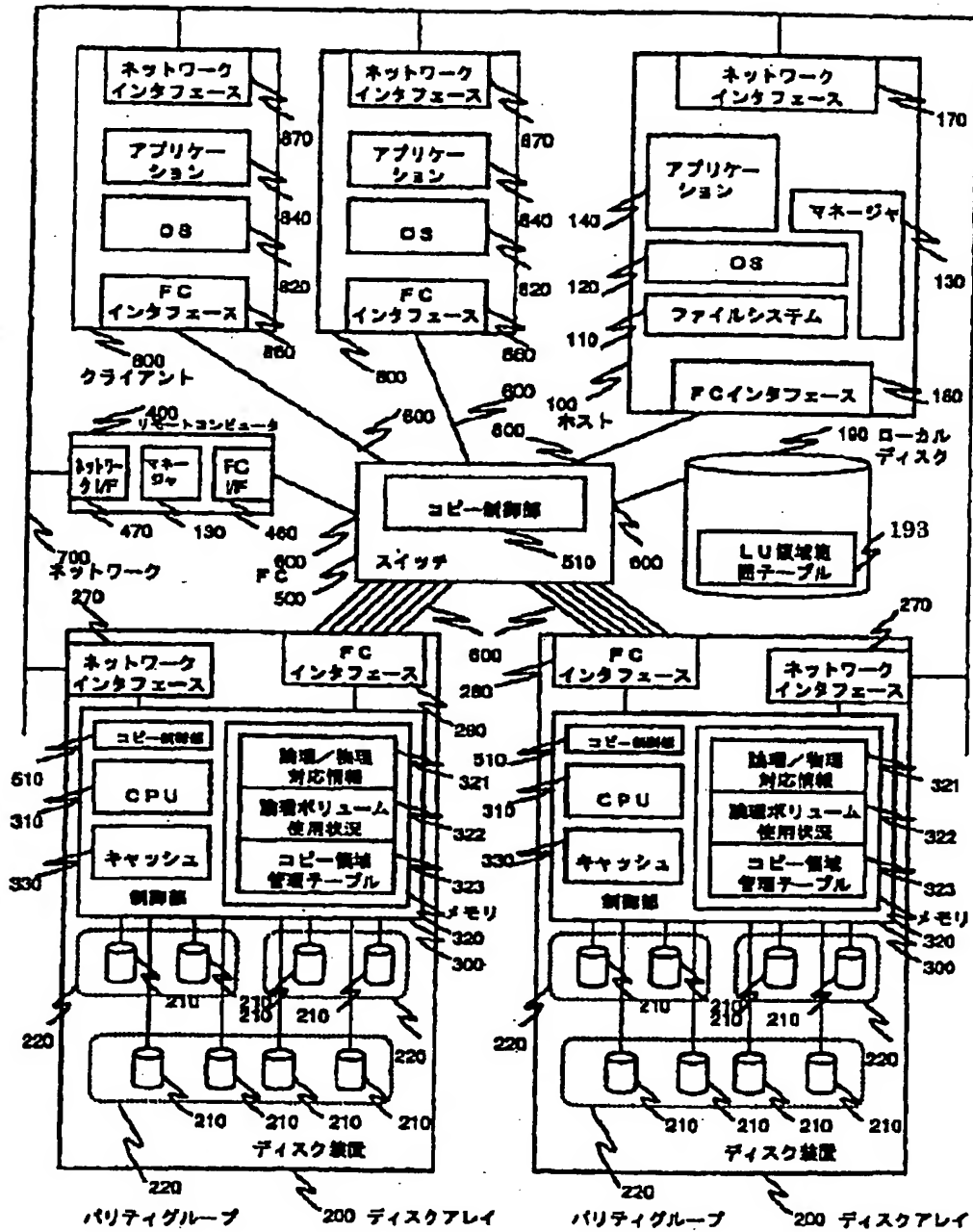
【図28】

図28



【図29】

図29



【図30】

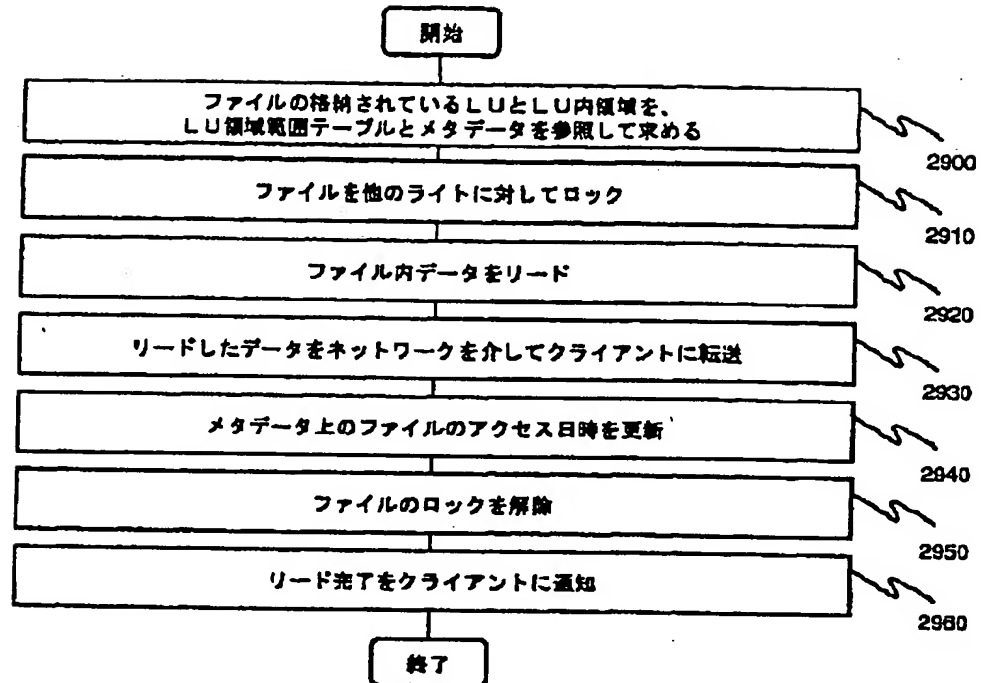
図30

3510

使用種別	領域内アドレス	LU番号			LU内アドレス
		ディスクアレイ 番号	ID	LUN	
FC経由データ転送 によるファイル共有	0~999	0	2	3	0~999
	1000~2999	1	4	5	0~999
	3000~3999	1	5	2	0~999
ネットワーク経由 データ転送による ファイル共有	0~999	0	2	4	0~999
	1000~2999	1	4	6	0~999
FC経由データ転送 によるLU共有	0~999	1	1	1	0~999
ネットワーク経由 データ転送による LU共有	0~999	1	2	2	0~999
未使用領域	—	0	0	0	0~999
	—	0	0	1	0~999

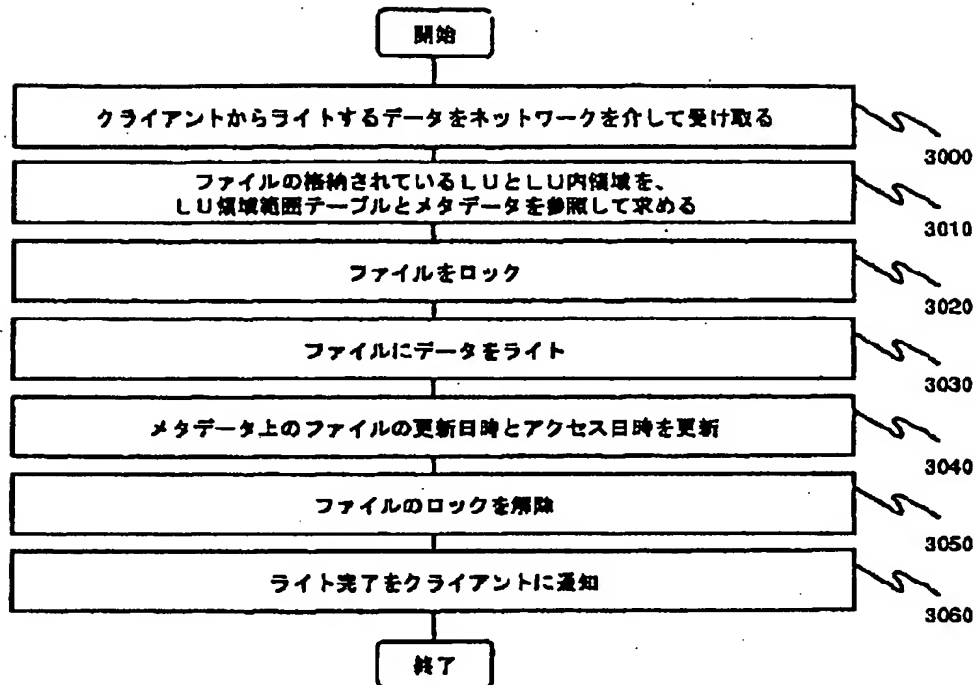
【図31】

図31



〔図32〕

図32



フロントページの続き

(72)発明者 江口 賢哲
神奈川県川崎市麻生区王禅寺1099番地 株
式会社日立製作所システム開発研究所内

(72)発明者 荒井 弘治
神奈川県小田原市国府津2880番地 株式会
社日立製作所ストレージシステム事業部内
Fターム(参考) 5B065 CA30 CH18